UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Estimación de personalidad basado en análisis de movimientos

Máster Universitario en Ingeniería Informática

Autor: Cañar Gutiérrez, Álex René

Tutor(es): Martínez Muñoz, Gonzalo – Pulido Cañabate, Estrella
Departamento de ingeniería informática

FECHA: Septiembre, 2019

Estimación de personalidad basado en análisis de movimientos

AUTOR: Álex René Cañar Gutiérrez
TUTOR(ES): Martínez Muñoz, Gonzalo – Pulido Cañabate, Estrella

Dpto. Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Septiembre de 2019

Resumen

El análisis de las personas según su comportamiento ha sido siempre algo de gran interés y, por tanto, objeto de múltiples estudios en el campo de la psicología.

Por otro lado, distintos sistemas de aprendizaje automático se están utilizando para hacer estimaciones futuras a partir del comportamiento de las personas. Asimismo, otros estudios más recientes además identifican qué tipo de movimientos están relacionados con determinadas características de la personalidad.

Este trabajo fin de máster tiene como fin construir un sistema de aprendizaje automático para detectar distintas facetas de la personalidad a partir de los movimientos de los sujetos en una entrevista y analizar los posibles patrones o factores que ayuden a caracterizar esa personalidad. Para llevar a cabo este estudio se parte de datos obtenidos en entrevistas mediante un dispositivo Kinect. Este dispositivo permite monitorizar el movimiento en distintas partes del cuerpo.

En primer lugar, se realiza un preprocesado de toda la información obtenida en crudo, identificando posibles atributos para posteriormente aplicar algoritmos de aprendizaje automático que ayuden a identificar la personalidad de los sujetos. Además, se calculan nuevos atributos más complejos, con este mismo fin. A continuación, se realiza un análisis de los resultados con el objetivo de filtrar los atributos que aporten más información acerca de la personalidad del sujeto.

Finalmente, se recopilan todos los resultados obtenidos de todas las pruebas realizadas y se extraen los atributos que mejores resultados hayan mostrado y se extraen una serie de conclusiones.

Palabras clave

Aprendizaje automático, Características de la personalidad, Kinect, Spark, Regresión lineal, Regresión logística, Random Forest Regressor, Random Forest Classifier, Validación cruzada, Error cuadrático medio, K-means, Clustering jerárquico

Abstract

People analysis based on behaviour has always been something of great interest, which is why so many psychology studies have been made about this topic. Many different machine learning systems are being used in order to make predictions based on people's behaviour. In addition, other recent studies also identify the type of movement which are related to certain personality features.

This Master's Thesis' purpose is to build a machine learning system able to detect different parts of human personality based on people's movements during an interview and analyze possible patterns or decisive factors that help identify said personality. In order to achieve that purpose, this project starts with data retrieval during the interviews through/with help of a device known as Kinect. This device allows monitoring the movement of different parts of the human body.

Firstly, a data pre-process is made by trying to identify possible attributes in order to apply machine learning algorithms, which help identify people's personality. Furthermore, more complex attributes are calculated to use them in the same way previously mentioned. Next, an analysis of the result is made in order to filter the attributes that give more information about people's personality.

Finally, the results from all tests are gathered and the attributes with best scores are extracted, then, some conclusions are made.

Keywords

Machine Learning, Personality, Kinect, Spark, Linear Regression, Random Forest Regressor, Logistic Regression, Random Forest Classifier, Cross-validation, Mean Square Error (MSE), K-means, Hierarchy clustering

Agradecimientos

En primer lugar, me gustaría agradecer a todas las personas que me hay apoyado durante toda esta carrera, ya que sin ellos esto no habría sido posible. También, quería agradecer enormemente a los profesores que he tenido durante estos años y que tanto me han enseñado, ya que sin sus conocimientos este trabajo no hubiera sido posible.

En segundo lugar, me gustaría agradecer todo el apoyo aportado por mi tutores, Estrella y Gonzalo, ya que confiaron en mí, y porque sin su ayuda esto no habría sido posible.

Por último, quería agradecer el apoyo de toda mi familia y mis amigos, los cuales siempre han estado apoyándome para que saliera adelante. En especial a Lira, Fede, José Carlos y el departamento de Analytics, por sus consejos y su conocimiento aportado que fue muy útil en momentos clave.

Índice

1 INTRODUCCIÓN 15

1.1 Motivación 15
1.2 Objetivos y planteamiento 15
1.3 Estructura del trabajo 16

2 ESTADO DEL ARTE 19

2.1 Psicología 19
2.2 Sensores de movimiento 19
2.3 Computación a gran escala 20
2.4 Aprendizaje automático 21

3 DESARROLLO 23

3.1 Descripción del dataset 23
3.2 Preprocesamiento de los datos 28
3.3 Extracción de características 29
3.4 Descripción de la configuración de los algoritmos de aprendizaje automático empleados 31

4 ANÁLISIS EXPERIMENTAL 35

4.1 Protocolo experimental 35
4.2 Análisis estadístico 35
4.2.1 Categorización de variables objetivo 37
4.3 Resultados 41
4.3.1 Características simples 41
4.3.2 Características compuestas 43
4.3.3 Clustering 44
4.4 Discusión de resultados 46

5 CONCLUSIONES Y TRABAJO FUTURO 49

5.1 Conclusiones 49
5.2 Trabajo futuro 49

BIBLIOGRAFÍA 51

ANEXOS I

A Características simples y descripciones I
B Estadísticas variable apertura III
C Estadísticas variable cordialidad VII
D Estadísticas variable responsabilidad XI
E Estadísticas variable neuroticismo XV
F Estadísticas variable extraversión XVII
G Clustering XXIII

Índice de figuras

Figura 1.Partes del cuerpo monitorizadas 23
Figura 2. Ejemplo de fichero de datos de una persona 25
Figura 3. Histogramas de las variables objetivo 27
Figura 4. Fichero de intervalos de tiempo de una persona específica 28
Figura 5. Diagrama Entidad - Relación 29
Figura 6. Configuración de Random Forest Regressor 31
Figura 7. Configuración de Linear Regression 31
Figura 8. Configuración de Random Forest Classifier 32
Figura 9. Configuración de Logistic Regression 32
Figura 10. Configuración de Cross Validation 32
Figura 11. Puntos del cuerpo - correlación neuroticismo grupal e individual 36
Figura 12. Puntos del cuerpo - correlación extraversión grupal e individual 37
Figura 13. Histograma para el neuroticismo individual categorizado 38
Figura 14. K-Means para 3 clusters 39
Figura 15. Clustering jerárquico con el método Ward 40
Figura 16. Resultados de las características simples 43
Figura 17. Resultados características complejas 44
Figura 18. Random Forest Classifier – Neuroticismo 45
Figura 19. K-Means - Regresión logística con distancia euclídea entre manos 46
Figura 20. Clustering jerárquico - RandomForestClassifier con características simples 46
Figura 21. Puntos del cuerpo - correlación apertura grupal e individual III
Figura 22. Regresión Lineal (a) y Random Forest Regression (b) para apertura grupal IV
Figura 23. Regresión Lineal (a) y Random Forest Regressor (b) con los 6 puntos más correlacionados para apertura grupal IV
Figura 24. Regresión Lineal (a) y Random Forest Regression (b) para apertura individual V
Figura 25. Regresión Lineal (a) y Random Forest Regressor (b) con los 6 puntos más correlacionados para apertura individual V
Figura 26. Puntos del cuerpo - correlación cordialidad grupal e individual VII
Figura 27. Regresión Lineal (a) y Random Forest Regression (b) para cordialidad grupal VIII
Figura 28. Regresión Lineal (a) y Random Forest Regression (b) con los 6 puntos más correlacionados para cordialidad grupal VIII
Figura 29. Regresión Lineal (a) y Random Forest Regression (b) para cordialidad individual IX
Figura 30. Regresión Lineal (a) y Random Forest Regression (b) con los 6 puntos más correlacionados para cordialidad individual IX
Figura 31. Linear Regression (a) y Random Forest Regression (b) con intervalos con máxima y mínima movilidad para cordialidad individual X
Figura 32. Puntos del cuerpo - correlación responsabilidad grupal e individual XI
Figura 33. Regresión Lineal (a) y Random Forest Regression (b) para responsabilidad grupal XII
Figura 34. Regresión Lineal (a) y Random Forest Regression (b) con los 6 puntos más correlacionados para responsabilidad grupal XII
Figura 35. Regresión Lineal (a) y Random Forest Regression (b) para responsabilidad individual XIII
Figura 36. Regresión Lineal (a) y Random Forest Regression (b) con los 6 puntos más correlacionados para responsabilidad individual XIII
Figura 37. Matriz de correlación - Neuroticismo individual distancia entre manos XV
Figura 38. Matriz de correlación – Neuroticismo individual categorizado XVI
Figura 39. Matriz de correlación – Extraversión individual categorizada XVII
Figura 40. Matriz de correlación – Extraversión individual distancia entre manos XVIII
Figura 41. Matriz de correlación – Extraversión individual distancia entre cabeza y pies XVIII
Figura 42. Matriz de correlación – Extraversión individual distancia euclídea entre manos XIX
Figura 43. Matriz de correlación – Extraversión individual distancia entre manos para dos intervalos de tiempo XX
Figura 44. Regresión Lineal (a) y Random Forest Regression (b) para extraversión individual por intervalos XXI
Figura 45. Histograma para la extraversión individual categorizada XXI

Figura 46. Random Forest Classifier – Extraversión XXII
Figura 47. K-means con 4 clusters XXIII
Figura 48. K-means con 5 clusters XXIII
Figura 49. Dendograma clustering jerárquico método single XXIV
Figura 50. Dendograma clustering jerárquico método complete XXV
Figura 51. Dendograma clustering jerárquico método average XXVI
Figura 52. Dendograma clustering jerárquico método weighted XXVII
Figura 53. Dendograma clustering jerárquico método centroid XXVIII
Figura 54. Dendograma clustering jerárquico método median XXIX
Figura 55. K-Means - Regresión logística (a) y RandomForestClassifier (b) con distancia entre cabeza y pies XXX
Figura 56. K-Means - Regresión logística (a) y RandomForestClassifier (b) con características simples XXX
Figura 57. K-Means - Regresión logística (a) y RandomForestClassifier (b) con distancia entre manos XXXI
Figura 58. K-Means - RandomForestClassifier con distancia euclídea entre manos XXXI
Figura 59. Clustering jerárquico - Regresión logística (a) y RandomForestClassifier (b) con distancia entre manos XXXII
Figura 60. Clustering jerárquico - Regresión logística (a) y RandomForestClassifier (b) con distancia euclídea entre manos XXXII
Figura 61. Clustering jerárquico - Regresión logística (a) y RandomForestClassifier (b) con distancia entre cabeza y pies XXXIII
Figura 62. Clustering jerárquico - Regresión logística con características simples XXXIII

Índice de tablas

Tabla 1. Parte del cuerpo – identificador 24
Tabla 2. Estructura de los ficheros de cada persona 25
Tabla 3. Estructura del fichero de intervalos de tiempo 27
Tabla 4. Umbrales empleados para realizar la clasificación 37
Tabla 5. Centroides para K-means con 3 clusters 39
Tabla 6. Centroides para clustering jerárquico 41
Tabla 7. Características simples - descripción I

1 Introducción

1.1 Motivación

La realización de este trabajo proviene del gran interés que siempre ha suscitado el análisis de las personas, ya que esto puede ser tomado como punto de partida para poder alcanzar otros objetivos como la predicción del comportamiento humano, y todas las implicaciones que esto conllevaría. Es por ello por lo que hoy en día existen múltiples estudios basados en el comportamiento humano que han dado lugar a su uso en aplicaciones reales. Por lo tanto, este trabajo está motivado por el deseo de analizar a las personas según su comportamiento y extraer cierta información en forma de patrones que puedan caracterizar a cada una de ellas. Además, se construirá un sistema de aprendizaje automático con este fin: la detección de distintas facetas de la personalidad de las personas en base a los movimientos realizados durante una entrevista.

El análisis de dichas personas tiene como punto de partida un estudio previo, el cual consistía en realizar una entrevista a cada persona mientras se le grababa en audio y video, y, que fue realizado por el Grupo de Investigación en Piscología y Ciencias del Deporte, quienes se encargaron de recoger los datos de relativos a cada persona. Para ello, se emplea un dispositivo Kinect, el cual es capaz de monitorizar el movimiento de distintas partes del cuerpo humano y generar unos datos con un formato concreto, los cuales serán el punto de partida del trabajo.

Una vez se obtuvieron los datos se procede con el preprocesamiento de estos, en el cual se verifica que no falte ningún dato, que todos los datos cumplan un mismo formato y que, por tanto, no existan incoherencias en los mismos. A continuación, se realiza el correspondiente procesamiento de los datos en el cual se extraen una serie de características simples y complejas, las cuales se emplearán en la siguiente fase del trabajo. Para ello, se emplean una serie de librerías y una infraestructura que permita el manejo de grandes cantidades de datos.

A continuación, se emplean las respectivas técnicas de aprendizaje automático, sobre los datos obtenidos tras la fase del procesamiento con el objetivo de identificar posibles atributos que ayuden a definir y predecir el comportamiento de las personas. Por otra parte, se emplea otro enfoque que consiste en agrupar las distintas facetas de la personalidad en grupos más genéricos, por lo que, se emplean los datos obtenidos durante la fase del procesamiento con el objetivo de encontrar características que identifiquen estos grupos. Además, para cada enfoque se comparan los resultados obtenidos por cada algoritmo y características empleadas, con el objetivo de determinar los atributos que se acerquen más al objetivo mencionado anteriormente.

1.2 Objetivos y planteamiento

El objetivo principal es la predicción de distintas facetas de la personalidad a partir de secuencias de movimiento de distintas partes del cuerpo. Los objetivos secundarios son:

- Análisis del uso de los distintos atributos estadísticos simples obtenidos de las secuencias de movimiento.

- Identificación de secuencias de movimiento que predigan la personalidad de los sujetos.
- Simplificación de las facetas de personalidad en grupos mediante técnicas de clustering para la predicción del grupo al que pertenecen los sujetos.

En cuanto al plan de trabajo se realizarán los siguientes pasos:

- Preprocesamiento de los datos obtenidos por la Kinect para cada usuario. Se implementará un programa que permita la limpieza y preprocesado de los datos de movimientos de forma eficiente, con el obtenido de uniformizar los datos (mismo número de movimientos para cada usuario, comprobación de los valores de las distintas facetas de la personalidad para cada usuario, comprobación de la existencia de los ficheros que relacionan a los usuarios con sus intervalos de tiempo… etc) y evitar futuras incoherencias y errores.
- Procesamiento de los datos obtenidos y automatización. Se implementará un programa que lleve a cabo una extracción de características automáticamente a partir de los movimientos de los sujetos. Con este objetivo se podrán crear numerosos atributos que nos permitirán predecir distintas facetas de la personalidad. Además, para el enfoque de la predicción de las distintas facetas de la personalidad, se trabajará en dos líneas para la extracción de atributos que caractericen los movimientos: movimientos simples relacionados con el movimiento de una única parte del cuerpo; y movimientos complejos que relacionen el movimiento de dos o más partes de éste. Para el enfoque relacionado con la predicción de grupos de personalidad genéricos, se emplearán las mismas líneas de trabajo explicadas anteriormente con la diferencia de que las variables objetivo serán distintas a las del anterior enfoque.
- Aplicación de diversos algoritmos de aprendizaje automático y comparativa de los resultados obtenidos, así como análisis de las variables más representativas para el estudio.
- Conclusiones en base a los resultados obtenidos anteriormente.

1.3 Estructura del trabajo

La memoria consta de los siguientes capítulos:

- **Estado del arte**

En este capítulo se analiza el estado de la tecnología actual en relación con nuestro proyecto. Se realiza una descripción de las distintas facetas de la personalidad, las cuales representan a las variables objetivo que se intentan predecir en este trabajo. Además, se describen las tecnologías a emplear junto con sus características más destacadas y su relación e importancia en este trabajo. Finalmente, se detallan los algoritmos que se emplean para la extracción de posibles patrones de comportamiento, que es el objetivo de este trabajo.

- **Desarrollo**

En este capítulo se explica la metodología de trabajo que se ha seguido, así como las partes con mayor relevancia. En primer lugar, se realiza una descripción detallada de los datos iniciales que se toman como punto de partida. Posteriormente, se detallan las características simples y complejas que se emplean y su relevancia en este trabajo. También, se realiza una descripción de los algoritmos que se emplean junto con la configuración empleada para cada uno de ellos. Por último, se explican los métodos empleados para la comparación de los resultados obtenidos para cada enfoque.

- **Análisis experimental**

En este capítulo se muestran los resultados de todas las pruebas realizadas. Primeramente, se presenta los resultados obtenidos al realizar un análisis inicial de los datos. A continuación, se detallan las distintas pruebas realizadas sobre cada una de las facetas de la personalidad. Además, se realiza una valoración de estos resultados. Finalmente, se exponen los resultados obtenidos para el segundo enfoque de este trabajo.

- **Conclusiones y trabajos futuros**

En este capítulo se exponen las conclusiones obtenidas en base a los resultados expuestos anteriormente. Para cada enfoque se analizan los resultados obtenidos y se presentan una serie de conclusiones en base a ello. Además, se presentan futuros trabajos que podrían derivarse de éste, así como, la justificación de su relación con este trabajo y sus posibles aportaciones al mismo.

2 Estado del arte

En este capítulo se exponen tanto los elementos conceptuales como la tecnología que conforman este trabajo. En la sección 2.1, se introducen las distintas dimensiones de la personalidad humana. En la sección 2.2 se introducen los sensores de movimiento y su relevancia con este trabajo. En la sección 2.3 se introduce la computación a gran escala y su relación e importancia en el trabajo. Finalmente, en la sección 2.4 se describe el aprendizaje automático y su importancia para con el trabajo.

2.1 Psicología

La psicología es una ciencia cuyo objetivo es el estudio y entendimiento del comportamiento humano [1]. El análisis de las personas a través de su comportamiento ha supuesto siempre un punto de gran interés y, por tanto, objeto de múltiples estudios en el campo de la psicología. Algunos ejemplos de estos estudios son el análisis del comportamiento en redes sociales de potenciales empleados para determinar su posible contratación o no [2], el análisis del comportamiento se utiliza también para determinar la libertad condicional de personas que están siendo procesadas [3], y también, diversos cuidados preventivos que se pueden proponer según los perfiles de los pacientes [4].

Por otro lado, existen múltiples estudios teóricos que relacionan los movimientos de una persona con la personalidad de ésta [5]. Otros estudios más recientes también identifican qué tipo de movimientos están relacionados con determinadas características de personalidad como aparece en el estudio de Kleinsmith y Bianchi-Berhouze [6]. De la misma manera el EPI (Eysenck Personality Inventory) tiene como objetivo la evaluación de dos dimensiones de la personalidad [7]:

- El neuroticismo, entendido como la reacción exagerada de tipo emocional y la predisposición a la depresión neurótica en situaciones de estrés [8].
- La extraversión, la cual, al contrario de la introversión [9], se refiere a las tendencias impulsivas hacia la exteriorización y la no inhibición de una persona [10].

Además, se analizan otras dimensiones de la personalidad como son:

- La responsabilidad, definida como un compromiso, una obligación o un deber, ya sea con los demás como con uno mismo [11].
- La cordialidad, definida como la amabilidad de una persona y su capacidad para establecer vínculos con otras personas [12].
- La apertura a la experiencia, definida por la capacidad de una persona para tener originalidad, creatividad, y entusiasmo para explorar nuevos caminos [13].

2.2 Sensores de movimiento

Los estudios de personalidad son de gran interés. Es por ello por lo que son necesarias herramientas que permitan cuantificar las diferentes áreas de estudio en datos concretos de

los que se puedan obtener información. En este aspecto, se presentan los sensores de movimiento, lo cuales permiten hacer una traducción de los movimientos de las personas durante el estudio con datos concretos con lo que se pueden aplicar diferentes técnicas para extraer información y conclusiones.

Uno de los dispositivos más empleados es la conocida como "Kinect". Este dispositivo desarrollado por Microsoft [14, 15] desde el año 2009, permite monitorizar distintas partes del cuerpo, así como los movimientos de éstas. Para ello, este dispositivo está formado por tres grandes componentes a nivel de hardware como son: una cámara de video, un sensor de profundidad y un micrófono [16]. A partir de estos elementos, el dispositivo es capaz de monitorizar 48 puntos diferentes del cuerpo humano a una velocidad de 30 frames por segundo. En cuanto al software, este dispositivo emplea un SDK, el Azure Kinect DK [17, 18], el cual permite manipular los datos provenientes del hardware.

Las aplicaciones de este dispositivo, más allá de su uso como dispositivo de juego, son abundantes y abarcan diferentes áreas, como pueden ser la salud, el comercio, la logística y la robótica [18].

2.3 Computación a gran escala

Con el desarrollo de las nuevas tecnologías, la cantidad de información disponible ha crecido de forma exponencial, por lo que surge, en paralelo, la necesidad de crear tecnologías que permitan gestionar y trabajar con cantidad de información. Una de las fuentes que alberga una gran cantidad de información es la grabación de video, a partir de la cual surgen otros dispositivos como la Kinect. En este sentido y, con el objetivo de dar soporte a esta necesidad, han surgido a lo largo de los últimos años una serie de tecnologías tanto software como hardware.

En la parte de hardware, han surgido múltiples empresas que han desarrollado su propia solución para enfrentar este problema, la cual consiste tanto en la creación de grandes centros de datos que albergan una gran cantidad de ordenadores, así como la creación de herramientas software que faciliten el uso de esos centros de datos de forma transparente al usuario. Entre las empresas más destacadas están Amazon [19], Microsoft [20], Databricks [21]. La mayoría de ellas nos proporcionan una prueba gratuita con la que poder desarrollar ciertas aplicaciones y comprobar la infraestructura, así como su eficiencia y rendimiento.

En cuanto a la parte del software, cabe destacar de entre las tecnologías de procesado de grandes cantidades de datos, Spark. Además, entre las principales características de Apache Spark en comparación con otras plataformas destacan su simplicidad y transparencia al usuario en cuanto al gestión de los recursos, así como su rapidez y concepción de plataforma de propósito general.

Spark tiene a Apache Hadoop [22] como una de sus bases. Hadoop fue una de las primeras tecnologías de procesado de datos masivos que se empleó. Entre sus características destacan su software de código libre, su alta escalabilidad y el empleo del paradigma Map/Reduce, el cual consiste en que la aplicación se divide en pequeños trozos, los cuales pueden ser ejecutados en cualquier elemento del cluster.

Apache Spark se diferencia de Apache Hadoop en que añade una gran variedad de mejoras como la velocidad, hasta 100 veces, ejecución interactiva y en streaming. También, el desarrollo de varias APIs en varios lenguajes, siendo la simplicidad de cada una de ellas, su principal característica [23].

Uno de los módulos más destacados de Spark es Spark SQL [24], el cual permite la ejecución de consultas SQL y, por tanto, el trabajo con datos estructurados. Esta información relativa a la estructura de datos junto con la información extra acerca de la operación a ejecutar es empleada por Spark, de forma interna, para realizar optimizaciones extra. Por otra parte, este módulo implementa una unificación entre el motor de ejecución y en la computación del resultado, por lo que los usuarios pueden cambiar con facilidad entre los distintos lenguajes de programación disponibles, y elegir aquel que les proporcione una forma más natural e intuitiva de realizar una operación.

En cuanto a la arquitectura, Spark se puede ejecutar tanto en formato standalone, es decir, en un ordenador personal, como sobre un cluster de ordenadores, así como la nube de Amazon y Hadoop. También, cabe destacar que es muy versátil en cuanto al formato de los ficheros que acepta, ya que pueden ser tanto ficheros locales, como HDFS, como ficheros provenientes de una base de datos o de una fuente externa como Amazon S3.

2.4 Aprendizaje automático

El aprendizaje automático es una disciplina cuyo objetivo es el desarrollo de programas que sean capaces de generalizar patrones en el comportamiento a partir del análisis de un conjunto de datos [25,26,27]. Algunas aplicaciones del aprendizaje automático incluyen la seguridad informática, como la detección de fraude fiscal [28], el procesamiento de imágenes [29], motores de recomendación [30], etc.

Atendiendo a la naturaleza de los datos, podemos establecer dos clases de aprendizaje automático: el aprendizaje automático supervisado y el aprendizaje automático no supervisado [31].

El **aprendizaje automático supervisado** se caracteriza por el hecho de que existe un conocimiento previo sobre los elementos, los cuales tienen asignada una etiqueta. El objetivo es crear un modelo, el cual a partir de los datos de ejemplo etiquetados sea capaz de asignar una etiqueta de salida a un nuevo dato de entrada [32]. Dentro del aprendizaje supervisado se distinguen principalmente dos tipos de tareas dependiendo del tipo de etiqueta que tienen los ejemplos. Hablaremos de clasificación cuando la etiqueta es categórica y de regresión cuando la etiqueta es numérica. A continuación, se describen algunos de los algoritmos más empleados:

- Arboles de decisión. En este algoritmo, se definen un conjunto de reglas para decidir qué clasificación es la más adecuada para cada elemento en base a sus atributos [33].
- Regresión lineal. Esta técnica de regresión tiene como objetivo encontrar un hiperplano que represente de forma óptima al conjunto total de puntos [34]. Generalmente este algoritmo se ajusta minimizando la suma de los cuadrados de las distancias verticales entre cada punto y la recta en el conjunto de entrenamiento [35].

- Regresión logística. Este algoritmo tiene como base la probabilidad de que ocurra un evento en base a otros factores [36]. Por lo tanto, este algoritmo consiste en obtener una función logística de las variables independientes que permita identificar el grupo o categoría a la que pertenecen los individuos.
- Random Forest. Este algoritmo se basa en la combinación de múltiples arboles de decisión con el fin de determinar el resultado final [37]. Para ello se emplean árboles aleatorios y muestras Bootstrap para entrenamiento de estos. Para la regresión (Random Forest Regression) se puede observar que el resultado final es la media de las salidas de las funciones simples, que en este caso son arboles de decisión

$$g(x) = \frac{f_0(x) + f_1(x) + f_2(x) + \cdots}{n} \tag{1}$$

Para la clasificación (Random Forest Classifier) se emplea un conjunto de árboles los cuales emiten un resultado cada uno siendo el resultado final aquel que fue obtenido por mayoría.

El **aprendizaje automático no supervisado** se caracteriza por la ausencia de un conocimiento previo, es decir, los datos no tienen asociada ninguna etiqueta. En general, este tipo de aprendizaje automático trata a las entradas como un conjunto de variables aleatorias. En este tipo de aprendizaje, se pueden establecer dos tipos de problemas: clustering y asociación.

En clustering se intenta encontrar patrones en los datos que conformen grupos independientes [31,32]. Los algoritmos más destacados son K-means [38] y clustering jerárquico [39]. K-means es un algoritmo por particiones y se caracteriza porque necesita conocer de ante mano el número de grupos a formar. Cada grupo o cluster formado está identificado por su centroide (centro geométrico del cluster). Clustering jerárquico es un algoritmo, el cual en cada fase del proceso genera una serie de grupos. En cada uno de los grupos, sus elementos están más relacionados entre sí que con los elementos de los otros grupos. Finalmente, el algoritmo encuentra el número de clusters que realizan una agrupación optima.

En asociación se intenta averiguar las reglas que describen a la mayor cantidad de datos posibles. Uno de los algoritmos más destacados es Apriori [40].

3 Desarrollo

En este capítulo se exponen las diferentes fases que se ha seguido en el desarrollo del trabajo. En la sección 3.1 se describen los datos con los que se van a trabajar. En la sección 3.2 se describen los atributos y características extraídas a partir de los datos iniciales. Finalmente, en la sección 3.3 se detallan los métodos de aprendizaje automático supervisado empleados.

3.1 Descripción del dataset

El conjunto de datos que se ha empleado para realizar este trabajo está basado un conjunto de entrevistas, realizadas por el Grupo de Investigación en Piscología y Ciencias del Deporte, en las cuales se realizaban a cada persona de forma individual una serie de cuestiones, con el objetivo tanto de recoger información relativa a la personalidad de cada persona como de obtener los movimientos realizados durante las mismas. Para ello, los 69 sujetos, los cuales eran estudiantes de psicología, eran grabados en audio y video, empleando una Kinect la cual, tal y como se menciona en el capítulo anterior, es un dispositivo capaz de detectar y almacenar el movimiento de las distintas partes de cuerpo y, estos datos, por tanto, conforman el conjunto de datos de partida. En la Figura 1, se pueden observar las diferentes partes del cuerpo humano sobre las que se han realizado estas medidas.

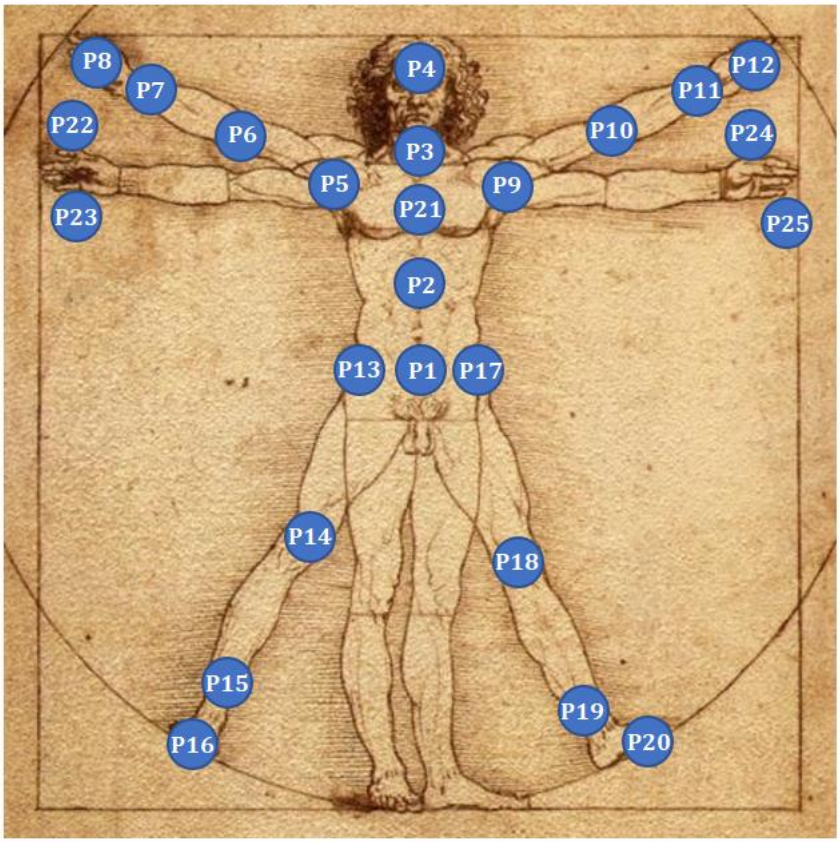


Figura 1.Partes del cuerpo monitorizadas

En la Tabla 1, se muestra un resumen de la Figura 1 que describe la relación entre las partes del cuerpo y su identificador asociado.

Parte del cuerpo	Identificador
Base Columna Vertebral	P1
Medio Columna Vertebral	P2
Cuello	P3
Cabeza	P4
Hombro Derecho	P5
Codo Derecho	P6
Muñeca Derecha	P7
Mano Derecha	P8
Hombro Izquierdo	P9
Codo Izquierdo	P10
Muñeca Izquierda	P11
Mano Izquierda	P12
Cadera Derecha	P13
Rodilla Derecha	P14
Tobillo Derecho	P15
Pie Derecho	P16
Cadera Izquierda	P17
Rodilla Izquierda	P18
Tobillo Izquierdo	P19
Pie Izquierdo	P20
Alto Columna Vertebral	P21
Dedo Pulgar Derecho	P22
Dedo Índice Derecho	P23
Dedo Pulgar Izquierdo	P24
Dedo Índice Izquierdo	P25

Tabla 1. Parte del cuerpo – identificador

Para cada uno de los puntos del cuerpo especificados anteriormente se recogen sus posiciones en los ejes:

- X: señala el eje horizontal.
- Y: señala el eje vertical.
- Z: señala el eje de profundidad.

Los puntos se recogen en metros con respecto al objetivo de la cámara, que es el punto X=0, Y=0, Z=0. Estos datos se recogen como una secuencia de video. La grabación suele tener una duración aproximada de 20 minutos, con entre 18 y 30 fotogramas por segundo. Todos estos datos han sido proporcionados en ficheros de texto plano.

Estos archivos están compuestos de un numero variable de filas donde cada fila representa un fotograma. El número de frames (filas del fichero) depende de la extensión de la grabación, pero su extensión es en cualquier caso de 10.000 líneas. Cada una de las filas incluye las posiciones xyz de cada una de las articulaciones descritas en la Tabla 1, así como

la información temporal del registro. En la Tabla 2 se muestran los campos de cada línea del fichero.

Columna	Descripción
C1	Hora del reloj del ordenador
C2	Minuto del reloj del ordenador
C3	Segundo del reloj del ordenador
C4	Milisegundo del reloj del ordenador
C5	Hora del reloj del ordenador en el video
C6	Minuto del reloj del ordenador en el video
C7	Segundo del reloj del ordenador en el video
C8	Milisegundo del reloj del ordenador en el video
C9 – C83	Posición X, Y, Z de los puntos 1 a 25
C84 – C108	Exactitud con la que se estiman los puntos 1 a 25

Tabla 2. Estructura de los ficheros de cada persona

En la Figura 2, se muestra un ejemplo con las primeras líneas de uno de los ficheros de salida de la Kinect.

```
1  10 17 54 942 001 005 051 952 0.01562 0.01706 1.87443 0.02125 0.31181 1.85517 0.02655 0.59416 1.82425 0.03066 0.74778 1.80810 -0.14173 0.48012 1.84023 -0.19720 0.21573
   1.83924 -0.20273 0.00641 1.74616 -0.19415 -0.05023 1.71270 0.19923 0.47799 1.81662 0.26606 0.22254 1.84103 0.27211 0.02502 1.72336 0.25514 0.00915 1.70946 -0.06289 0.01725
   1.84442 -0.31355 -0.09844 1.65030 -0.10499 -0.30808 1.65179 0.04602 -0.31403 1.66002 0.09356 0.01622 1.83459 0.25320 -0.15351 1.58339 0.24740 -0.44704 1.58358 0.16681
   -0.34454 1.64932 0.02528 0.52530 1.83399 -0.19517 -0.11417 1.66690 -0.21994 -0.00916 1.69569 0.23095 -0.02803 1.70363 0.27987 0.01304 1.68886 1 1 1 1 1 2 2 2 1 2 2 2 1 2 2
   2 1 2 2 1 1 1 1 1 1
2  10 17 54 983 001 005 051 952 0.01562 0.01706 1.87443 0.02125 0.31181 1.85517 0.02655 0.59416 1.82425 0.03066 0.74778 1.80810 -0.14173 0.48012 1.84023 -0.19720 0.21573
   1.83924 -0.20273 0.00641 1.74616 -0.19415 -0.05023 1.71270 0.19923 0.47799 1.81662 0.26606 0.22254 1.84103 0.27211 0.02502 1.72336 0.25514 0.00915 1.70946 -0.06289 0.01725
   1.84442 -0.31355 -0.09844 1.65030 -0.10499 -0.30808 1.65179 0.04602 -0.31403 1.66002 0.09356 0.01622 1.83459 0.25320 -0.15351 1.58339 0.24740 -0.44704 1.58358 0.16681
   -0.34454 1.64932 0.02528 0.52530 1.83399 -0.19517 -0.11417 1.66690 -0.21994 -0.00916 1.69569 0.23095 -0.02803 1.70363 0.27987 0.01304 1.68886 1 1 1 1 1 2 2 2 1 2 2 2 1 2 2
   2 1 2 2 1 1 1 1 1 1
3  10 17 55 280 001 005 052 286 0.02238 0.03013 1.86100 0.02370 0.31697 1.84856 0.02641 0.59407 1.82306 0.03090 0.74709 1.80636 -0.14149 0.48009 1.83750 -0.19397 0.22025
   1.82647 -0.16862 0.02133 1.69882 -0.14941 -0.04250 1.65594 0.19913 0.47780 1.81594 0.26583 0.22227 1.84068 0.27233 0.02382 1.72785 0.25435 0.00695 1.71527 -0.05611 0.03043
   1.83032 -0.09466 0.00212 1.32779 -0.04754 -0.13088 1.73704 0.06967 -0.18899 1.68888 0.10004 0.02874 1.82266 0.07988 -0.00375 1.30920 0.15581 -0.24990 1.72739 0.13216
   -0.34039 1.69115 0.02543 0.52595 1.83181 -0.12720 -0.10549 1.61837 -0.11908 -0.03060 1.67871 0.23132 -0.02931 1.71169 0.28241 0.00647 1.70485 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1
   2 1 1 1 1 1 1 1 1 1
4  10 17 55 364 001 005 052 286 0.02238 0.03013 1.86100 0.02370 0.31697 1.84856 0.02641 0.59407 1.82306 0.03090 0.74709 1.80636 -0.14149 0.48009 1.83750 -0.19397 0.22025
   1.82647 -0.16862 0.02133 1.69882 -0.14941 -0.04250 1.65594 0.19913 0.47780 1.81594 0.26583 0.22227 1.84068 0.27233 0.02382 1.72785 0.25435 0.00695 1.71527 -0.05611 0.03043
   1.83032 -0.09466 0.00212 1.32779 -0.04754 -0.13088 1.73704 0.06967 -0.18899 1.68888 0.10004 0.02874 1.82266 0.07988 -0.00375 1.30920 0.15581 -0.24990 1.72739 0.13216
   -0.34039 1.69115 0.02543 0.52595 1.83181 -0.12720 -0.10549 1.61837 -0.11908 -0.03060 1.67871 0.23132 -0.02931 1.71169 0.28241 0.00647 1.70485 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1
   2 1 1 1 1 1 1 1 1 1
```

Figura 2. Ejemplo de fichero de datos de una persona

Tras la realización de las entrevistas, el Grupo de Investigación en Piscología y Ciencias del Deporte evaluó la personalidad de los participantes en cinco dimensiones**: extraversión, apertura, cordialidad, neuroticismo y responsabilidad**. Estas evaluaciones se realizaron por parte de los propios participantes en sesiones individuales y grupales y, por allegados y expertos. En el presente trabajo se limitarán a las evaluaciones de los propios participantes. Posteriormente se generó un fichero en formato Excel con el siguiente formato:

- id_orden. Identificador del orden en el que se realizaron las grabaciones
- id_usu. Identificador de usuario que se corresponde con el número del archivo de datos de la Kinect.
- consent. 1 ha firmado el consentimiento para participar; 0 no ha firmado.
- consentimagen. 1 ha firmado el consentimiento para que enseñemos su grabación en actos de difusión científica; 0 no ha firmado.
- Genero. 1 femenino; 0 masculino.
- Edad. Edad de la persona, en años.
- total_NEUR_grupal. Puntuación en la dimensión de neuroticismo, obtenida en el cuestionario NEO FFI [41] administrado en sesión grupal al inicio del experimento.

- total_EXTR_grupal. Puntuación en la dimensión de extraversión, obtenida en el cuestionario NEO FFI administrado en sesión grupal al inicio del experimento.
- total_APER_grupal. Puntuación en la dimensión de apertura a la experiencia, obtenida en el cuestionario NEO FFI administrado en sesión grupal al inicio del experimento.
- total_CORD_grupal. Puntuación en la dimensión de cordialidad, obtenida en el cuestionario NEO FFI administrado en sesión grupal al inicio del experimento.
- total_RESP_grupal. Puntuación en la dimensión de responsabilidad, obtenida en el cuestionario NEO FFI administrado en sesión grupal al inicio del experimento.
- total_NEUR_indiv. Puntuación en la dimensión de neuroticismo, obtenida en el mismo cuestionario NEO FFI, pero administrado en sesión individual entre 1 y 2 semanas después.
- total_EXTR_indiv. Puntuación en la dimensión de extraversión, obtenida en el mismo cuestionario NEO FFI, pero administrado en sesión individual entre 1 y 2 semanas después.
- total_APER_indiv. Puntuación en la dimensión de apertura a la experiencia, obtenida en el mismo cuestionario NEO FFI, pero administrado en sesión individual entre 1 y 2 semanas después.
- total_CORD_indiv. Puntuación en la dimensión de cordialidad, obtenida en el mismo cuestionario NEO FFI, pero administrado en sesión individual entre 1 y 2 semanas después.
- total_RESP_indiv. Puntuación en la dimensión de responsabilidad, obtenida en el mismo cuestionario NEO FFI, pero administrado en sesión individual entre 1 y 2 semanas después.

En el fichero mencionado con anterioridad a cada individuo se le asignan 5 valores de personalidad. Estos valores representan las variables objetivo de este trabajo, las cuales se intentan predecir a partir de las secuencias de movimiento de cada individuo con algoritmos de aprendizaje automático. Además, estos valores de la personalidad son variables continuas con un rango entre 0 y 100. En la Figura 3 se muestran los histogramas relativos a dichas variables. Como se puede observar, los valores de las variables objetivo están muy agrupados en torno al rango [5,50] con picos en torno al valor 32 en la apertura grupal o en la cordialidad individual. Dado que las entrevistas fueron realizadas a estudiantes de psicología es razonable pensar que exista cierta similitud entre los individuos.

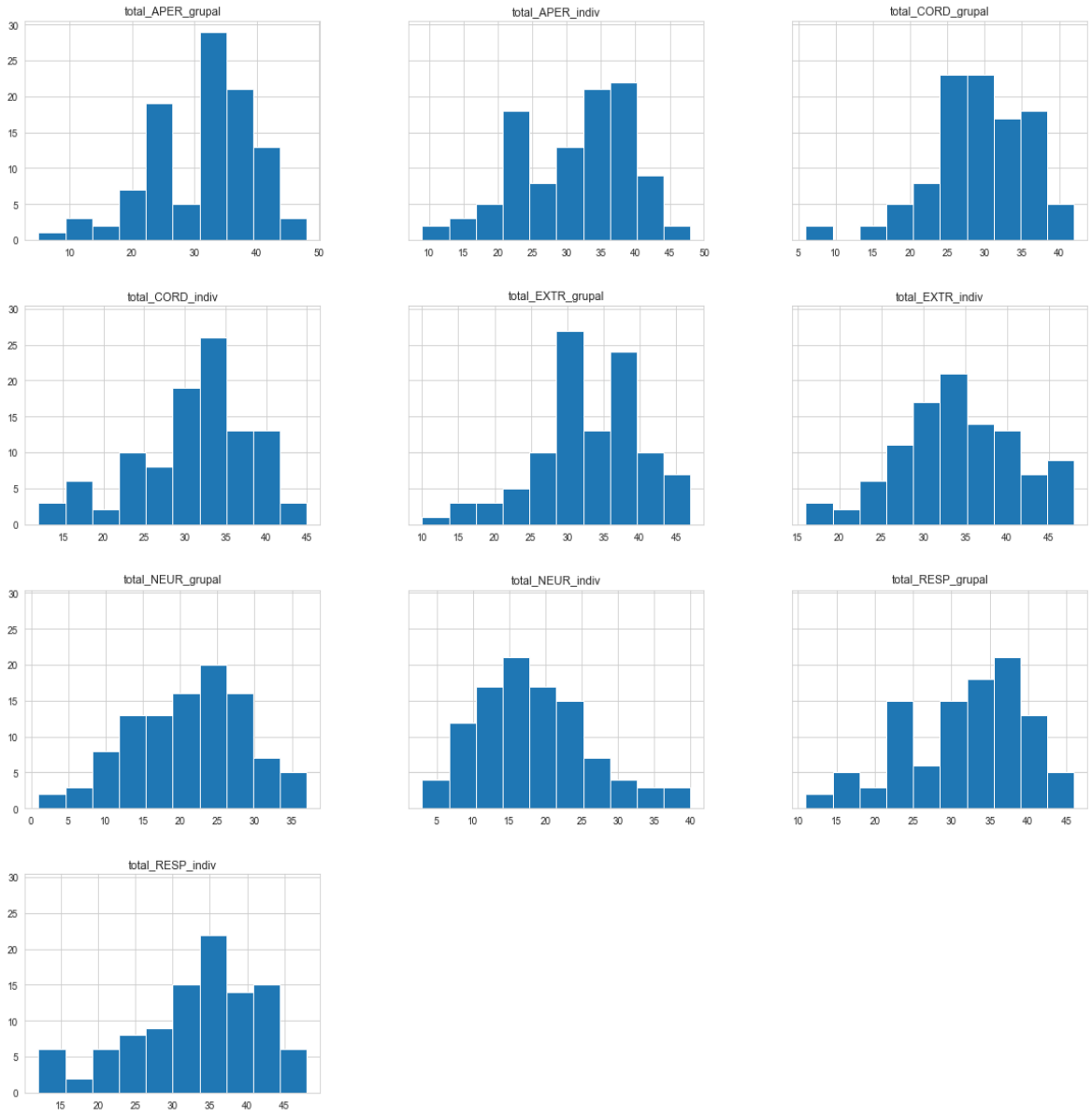


Figura 3. Histogramas de las variables objetivo

Como se ha comentado anteriormente, las secuencias de movimiento obtenidas se corresponden con los movimientos de los individuos durante las entrevistas. Adicionalmente, se obtienen los intervalos de tiempos durante las entrevistas, en los cuales los participantes hablan. Estos intervalos de tiempo tienen una relevancia mayor que otras partes de la secuencia, ya que permiten omitir periodos en los cuales los individuos realicen movimientos ajenos a la entrevista. Además, es conveniente destacar que estos intervalos de tiempos toman como punto de referencia el video, y no la secuencia de movimiento, por lo que, pueden no estar sincronizados con los intervalos de tiempo de las secuencias de movimiento recibidas. Estos intervalos de tiempo se obtuvieron mediante un fichero cuya estructura es la que se muestra en la Tabla 3.

Columna 1	Columna 2	Columna 3
Tiempo de inicio (ns)	Tiempo de finalización (ns)	Etiqueta

Tabla 3. Estructura del fichero de intervalos de tiempo

En la Figura 4 se muestra un fichero de este tipo con la estructura mencionada en la Tabla 3, con los intervalos para una persona específica. Como se puede observar, en este caso tendríamos cuatro intervalos de tiempo, cada uno delimitado por un instante de inicio y un instante de finalización.

```
1  0 708392931 <START>
2  708392931 1359068610 <END>
3  1359068610 1404946425 <START>
4  1404946425 1564285504 <END>
5  1564285504 1654067895 <START>
6  1654067895 2026023514 <END>
```

Figura 4. Fichero de intervalos de tiempo de una persona específica

3.2 Preprocesamiento de los datos

Una vez se conoce la estructura de datos con la que se va a trabajar, se realiza un análisis previo, el cual se realiza mediante un programa en Python. Este programa se encarga de procesar los ficheros de movimiento transformándolos en objetos que puedan ser procesados por Spark. Para ello, el programa recibe como dato de entrada el fichero de datos con la secuencia de movimiento de un individuo, y mediante una serie de librerías, transforma este fichero en un objeto de tipo RDD, que es almacenado como un fichero parquet, que permite acelerar la carga de datos y las posteriores operaciones. Durante la ejecución de este programa se detectan distintos fallos en varios ficheros. Por ejemplo, en algunos ficheros faltaban medidas para el último fotograma. Para estos casos se decide omitir las medidas para este último fotograma.

Por otra parte, se realiza un programa en Python, que se encargar de procesar el fichero Excel que relaciona a las personas con sus etiquetas (descrito previamente en la Sección 3.1) transformándolo en un objeto compatible con Spark. Este programa recibe como entrada este fichero y mediante una serie de librerías, lo convierte en una tabla SQL. Esto se hace para simplificar el proceso de carga por parte de los programas posteriores que realizan el procesamiento de los datos. Previamente a esto, en una inspección visual se observa que faltan las etiquetas para el usuario 9, por lo que se elimina a este individuo del análisis.

Por otro lado, se realiza un programa en Python, que se encarga de procesar los ficheros que contienen los intervalos de tiempo donde el participante está hablando (ver Tabla 3), transformándolos en objetos compatibles con Spark. Al realizar esto, se observó que existían ficheros duplicados con diferente contenido, por lo que se opta por descartar los últimos.

Tras el procesado de estos ficheros se obtienen 3 tablas SQL que almacenan la información de los ficheros. En la Figura 5 se puede observar el diagrama entidad-relación de dichas tablas. En la tabla *personid*, se almacena la relación entre el nombre del fichero y el individuo al que hace referencia ese fichero. En la tabla *persontags* se almacenan los datos correspondientes a los usuarios y sus correspondientes etiquetas. Por último, en la tabla *personusersmatching* se almacena la relación entre el individuo y su identificador en los ficheros de intervalos de tiempo.

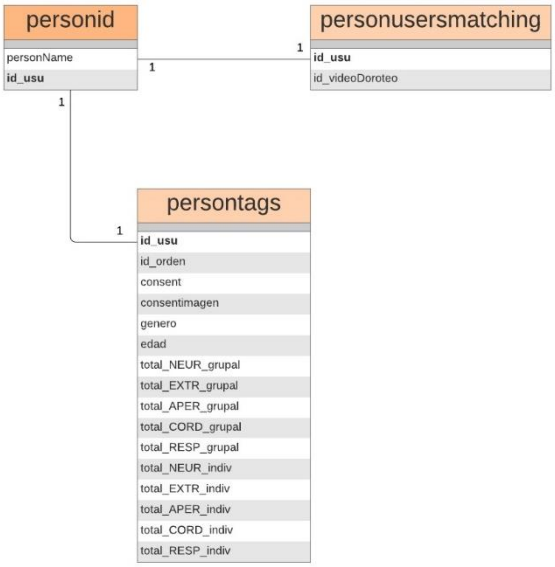


Figura 5. Diagrama Entidad - Relación

3.3 Extracción de características

Tras el preprocesamiento del conjunto de datos se procede a la extracción de atributos y características. Para ello, como se ha comentado anteriormente se dispone de las posiciones de cada parte del cuerpo en sus tres ejes (x, y, z) a lo largo de la entrevista. Esta información en si no tiene la suficiente relevancia ya que se trata de un conjunto de secuencias que varían entre sujetos y que no nos permiten extraer conclusiones aplicables a todos los sujetos. En cuanto a las características que se extrajeron, podemos clasificarlas en características simples y características compuestas.

Con características simples se hace referencia a aquellas características que se extraen de forma directa a partir de la secuencia de posiciones de cada parte del cuerpo. Se pueden ver las variables en más detalle en el anexo A. Para cada coordenada y articulación se utilizarán los siguientes atributos:

- Máximo: Máximo valor de la secuencia de la coordenada y articulación.
- Mínimo: Mínimo valor de la secuencia de la coordenada y articulación.
- Media: Valor medio de la secuencia de la coordenada y articulación.
- Desviación estándar: Dispersión de los valores de la secuencia de la coordenada y articulación.
- Suma de las diferencias con respecto a la media: Suma de las distintas diferencias, en valor absoluto, de cada valor de la secuencia de la coordenada y articulación con respecto a su media.
- Movilidad: Esta estadística informa acerca de la proporción de desviación típica del espectro de potencia [42]. Se calcula según la ecuación (2).

$$M_x = \sqrt{\frac{var(x'(t))}{var(x(t))}} \tag{2}$$

Donde, $var(x'(t))$ representa la varianza de la primera derivada de la señal analizada, mientras que $var\big(x(t)\big)$ representa la varianza de la señal analizada.

- Complejidad: Esta estadística analiza la similitud de la señal al compararla con una sinusoidal pura [45]. Se calcula mediante la ecuación (3).

$$C_x = \frac{M(x'(t))}{M(x(t))} \tag{3}$$

Donde, $M(x'(t))$ representa la movilidad de la primera derivada de la señal analizada mientras que $M(x(t))$ representa la movilidad de la señal analizada.

En el Anexo A podemos observar todos los identificadores de las características junto con su descripción. De ahora en adelante, se empleará la siguiente nomenclatura para la identificación en todas las características tanto simples como complejas:

- min[Característica]. Representa el mínimo valor para esa característica.
- max[Característica]. Representa el máximo valor para esa característica.
- avg[Característica]. Representa el valor medio de esa característica.
- stddev[Caractertistica]. Representa la desviación estándar para esa característica.
- sum[Característica]. Representa la suma de diferencias de cada valor con respecto a la media para esa característica.
- M[Característica]. Representa la movilidad para esa característica.
- C[Característica]. Representa la complejidad para esa característica.

Adicionalmente a las características simples, se han definido una serie de características compuestas con el objetivo de capturar acciones más complejas en los individuos. Con esto buscamos que alguna de esas acciones tenga una correlación con alguna de las variables objetivo (ver sección 3.1) y nos permita obtener buenas predicciones. Las características compuestas empleadas son las siguientes:

- Distancia entre las manos por eje: Esta característica consiste en calcular la distancia entre las manos en cada eje. Además, se ha optado por esta característica ya que se cree ciertas partes de la personalidad tienen una relación directa con cómo se mueven las manos. Su cálculo se obtiene de la diferencia en valor absoluto de cada eje de las manos (puntos 8 y 12 de la Tabla 1). Una vez obtenidas las secuencias con las distancias, se extraen las características simples de la secuencia. Esto es: máximo, mínimo, media, desviación típica, movilidad y complejidad.
- Distancia euclídea entre las manos: Esta característica consiste en calcular la distancia euclídea de las manos. El motivo por el cual se ha elegido esta característica es debido a su posible correlación directa entre esta y ciertas variables

objetivo a predecir. Para la obtención de esta característica, se calcula la distancia euclídea entre los puntos 8 y 12 (ver Tabla 1), entre los ejes x, y, z.

- Inclinación de la cabeza con respecto a los pies: Esta característica representa la distancia relativa entre la cabeza y los pies. Se cree que esta característica tiene una posible correlación con algunas de las variables objetivo, ya que las partes donde se enfoca son indicativas de alguna de las variables objetivo en este contexto. Para calcular esta característica se obtiene la posición de los pies (P16 y P20, ver Tabla 1) en el eje Z, y se calcula la diferencia en valor absoluto con la posición de la cabeza (P4, ver Tabla 1) en el eje Z.
- Distancia entre las manos para intervalos de tiempo específicos: Esta característica consiste en calcular la distancia entre las manos en cada eje para intervalos de tiempo de 1 minuto, con un desplazamiento de 5 segundos. El objetivo es calcular el intervalo de tiempo donde se obtiene el máximo valor de la variable movilidad en el eje X, y el intervalo de tiempo donde se obtiene el mínimo valor de la variable movilidad en el eje X. Posteriormente, se extraen las características simples de cada eje.

3.4 Descripción de la configuración de los algoritmos de aprendizaje automático empleados

Tras la obtención de las características y etiquetas asociadas a cada individuo, se procede a la aplicación de algunos algoritmos de aprendizaje automático, con el objetivo de obtener predicciones. Los valores de los parámetros, empleados en cada algoritmo, que se describen a continuación se corresponden con aquellos que mejores resultados han proporcionado. Por otra parte, se dispone de un conjunto de datos adicional para cada una de las características complejas. Cada conjunto de datos obtenido contiene 67 filas, una por individuo, que se han dividido en 17 filas (25%) para test y el resto para entrenamiento (75%). Los algoritmos que se emplean son los siguientes:

- Random Forest Regressor: Se emplea este algoritmo con una profundidad máxima en el árbol de 20, sin estado random inicial y con 100 árboles en cada bosque. En la Figura 6 se muestra la configuración descrita anteriormente.

```
RandomForestRegressor(max_depth=20, random_state=0,n_estimators=100)
```

Figura 6. Configuración de Random Forest Regressor

- Linear Regression: Se emplea este algoritmo con la configuración por defecto. En la Figura 7 se muestra dicha configuración.

```
LinearRegression()
```

Figura 7. Configuración de Linear Regression

- Random Forest Classifier: Se emplea este algoritmo con una profundidad máxima en el árbol de 20, sin estado random inicial y con 100 árboles en cada bosque. En la Figura 8 se muestra la configuración descrita anteriormente.

```
RandomForestClassifier(n_estimators=100, max_depth=20,random_state=0)
```

Figura 8. Configuración de Random Forest Classifier

- Logistic Regression: Se emplea este algoritmo con una fuerte regularización, además, se emplea *lbfgs* como algoritmo para resolver el problema y, se emplea la opción multinomial con 150 iteraciones como máximo, ya que con menos el algoritmo no converge. En la Figura 9 se muestra la configuración descrita anteriormente.

```
LogisticRegression(C=1e5, solver='lbfgs', multi_class='multinomial',max_iter=150)
```

Figura 9. Configuración de Logistic Regression

Adicionalmente, se han normalizado los atributos para que tengan media 0 y varianza 1. Para ello, se resta la media y se divide por la desviación estándar

$$X_{normalizado} = \frac{X - X_{media}}{X_{desviación\ estándar}} \tag{4}$$

Esta transformación es necesaria debido a que el rango de los datos es variable y muy dispar para algunas características, lo que hace que los algoritmos de regresión logística y regresión lineal no converjan correctamente. A los algoritmos RandomForestRegressor y RandomForestClassifier no les afecta esta transformación ya que los árboles generan divisiones equivales con los datos transformados o sin transformar.

Para la validación de los modelos se utiliza validación cruzada en todas las pruebas. Debido a la poca cantidad de datos se emplea una validación cruzada de 4 pliegues y se divide el conjunto de datos en un 75 por ciento para entrenamiento y el 25 restante para test. En la Figura 10 se muestra la configuración empleada.

```
KFold(n_splits=4)
```

Figura 10. Configuración de Cross Validation

En cuanto a la forma de comparar los resultados, se emplea el error cuadrático medio (MSE), ecuación (5). Además, como se ha mencionado con anterioridad, al emplear cross-validation se toma como referencia la media (AVG(MSE)) de las 4 pruebas realizadas.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - X'_i)^2 \tag{5}$$

Donde, X_i representa el valor real mientras que X'_i representa el valor predicho por el modelo.

En cuanto al objetivo de obtener varios clusters a partir de las variables objetivo, se emplean los algoritmos K-means y clustering jerárquico.

- El algoritmo K-means se emplea con un numero de clusters que varía en función de cada prueba, pero cuyo valor siempre está entre 3 y 5, se emplea como método de inicialización random.
- En el algoritmo de clustering jerárquico se emplea la configuración por defecto, y se emplean como métodos de unión de clusters, todos los disponibles en la librería, que son Single, Complete, Average, Weighted, Centroid, Median y Ward.

4 Análisis experimental

En este capítulo se presentan los resultados obtenidos en las diferentes pruebas realizadas. En la sección 4.1 se explica el procedimiento seguido para obtener los resultados. En la sección 4.2 se muestran un análisis estadístico inicial. En la sección 4.3 se muestran los resultados obtenidos. Finalmente, en la sección 4.4 se realiza una valoración de los resultados obtenidos.

4.1 Protocolo experimental

El proceso de obtención de los resultados consta de una primera fase en la cual se cargan los ficheros, en formato csv, obtenidos en el procesado de datos. Para ello, se añaden las librerías correspondientes y se realiza un primer análisis estadístico de cada fichero. Este análisis consiste en obtener las estadísticas tradicionales como la media y desviación típica. A continuación, se realiza un análisis más exhaustivo que implica la obtención de las correlaciones de las variables con respecto a la variable objetivo y la matriz de correlaciones. Las variables objetivo empleadas han sido las versiones grupal e individual de la apertura, la responsabilidad, la cordialidad, la extraversión y el neuroticismo.

En una segunda fase, se procede a la normalización de los datos, con el objetivo de obtener unos datos con media 0 y desviación típica de 1. Tras esto, se realiza la división de los datos en 4 pliegues (cross-validation) a través de la librería correspondiente de python, y se obtienen los índices de cada una de las muestras que se emplearán. Después, dependiendo del algoritmo que se emplee, se realiza la configuración de dicho algoritmo, tal y como se ha comentado en la sección 3.3. Para todas las variables objetivo se han realizado una serie de pruebas comunes. Estas pruebas consisten en para cada variable objetivo se emplean las características simples (descritas en la sección 3) junto con los algoritmos Linear Regression y Random Forest Regression. Otra prueba consiste en emplear los seis puntos del cuerpo más correlacionados con la variable objetivo, mostrados en la sección 4.2, y los algoritmos Linear Regression y Random Forest Regressor.

Posteriormente, se obtienen los resultados y se realiza el proceso de desnormalización de estos, con el objetivo de obtener la métrica del error cuadrático medio en la misma escala que la variable objetivo. A continuación, se obtiene la métrica mencionada anteriormente para cada uno de los 4 pliegues, se calcula le media de estos errores y se obtienen las gráficas correspondientes al algoritmo.

4.2 Análisis estadístico

En esta sección se muestran los resultados obtenidos al realizar un análisis estadístico inicial. Este análisis consiste en obtener las correlaciones de las características simples, ordenadas de forma descendente en relación con la variable objetivo.

En las figuras 11 y 12, se muestran las correlaciones para los puntos del cuerpo tanto para la variable objetivo neuroticismo grupal e individual, como la variable objetivo extraversión grupal e individual, ordenados de forma descendente. Para el resto de las variables, los resultados están en los anexos correspondientes. Como se puede comprobar, las

correlaciones mayores aparecen en las variables extraversión y neuroticismo. En general hay coincidencia entre los puntos más correlacionados en grupal e individual, sobre todo en neuroticismo. Además, se puede constatar que las correlaciones son bajas, por lo que no se esperan datos demasiado buenos. También, es evidente que los puntos 8 y 12 (mano derecha e izquierda) aparecen mucho, lo cual es indicativo de que tienen relevancia en las facetas de la personalidad. Las características simples movilidad y complejidad aparecen frecuentemente entre las más correlacionadas, esto puede ser debido a la naturaleza de los datos, ya que son secuencias temporales, y estas dos características tiene gran relación con ese tipo de datos.

total_NEUR_grupal	1.000000	total_NEUR_indiv	1.000000
CX4	0.286069	CY8	0.286273
CY8	0.280813	CX5	0.249347
CX8	0.236984	CX4	0.236066
CX9	0.230365	CX9	0.194886
CX5	0.222192	CX8	0.167625
minY12	0.177955	minY5	0.152356
SumDiffAvgY5	0.174394	MX12	0.145359
SumDiffAvgY9	0.173265	minY4	0.144636
stddevY12	0.160761	CX3	0.136447
SumDiffAvgZ5	0.158898	minY8	0.136048
SumDiffAvgZ3	0.147747	minY9	0.134819
SumDiffAvgY3	0.146655	CZ9	0.134059
SumDiffAvgZ4	0.138782	stddevY12	0.133229
stddevY8	0.134874	SumDiffAvgY12	0.123381
SumDiffAvgY12	0.128279	SumDiffAvgX8	0.118133
SumDiffAvgY4	0.127670	minY3	0.117894
minY8	0.127413	MZ12	0.111639
SumDiffAvgY8	0.124229	minY12	0.110373
SumDiffAvgZ9	0.124192	minZ3	0.108107
CZ9	0.100155	minZ4	0.107285
SumDiffAvgZ12	0.099054	minZ5	0.106946
CX3	0.090925	minX12	0.106322
stddevY9	0.081731	CZ3	0.100083
minX8	0.079581	minX3	0.094678
CY3	0.077262	minX8	0.093432
avgY8	0.073482		

Figura 11. Puntos del cuerpo - correlación neuroticismo grupal e individual

total_EXTR_grupal	1.000000	total_EXTR_indiv	1.000000
CZ3	0.362853	CZ8	0.310107
CZ8	0.304340	maxY5	0.261582
maxY5	0.210511	stddevZ8	0.258466
MX5	0.190922	stddevZ12	0.247395
maxY3	0.186812	maxY3	0.237679
MX9	0.181503	maxY4	0.233662
maxY4	0.173299	stddevZ9	0.222239
avgY8	0.168968	CX12	0.208504
MX3	0.160518	SumDiffAvgY12	0.202072
MX4	0.153830	maxZ9	0.200528
maxY9	0.145826	SumDiffAvgZ8	0.192390
CX3	0.142728	stddevZ5	0.192012
stddevY12	0.140277	stddevY5	0.191557
maxX5	0.127474	avgZ5	0.185410
CZ9	0.126449	maxY9	0.184004
CZ5	0.124104	maxZ5	0.183429
SumDiffAvgY12	0.121462	stddevZ3	0.181287
MX12	0.121283	stddevY9	0.179953
stddevX8	0.111973	stddevZ4	0.175539
SumDiffAvgX8	0.110287	stddevY3	0.173571
CX8	0.104451	avgZ4	0.172423
maxY12	0.095392	avgZ9	0.163590
maxX9	0.094664	maxZ4	0.162566
minY5	0.089764	stddevY12	0.162328
MZ3	0.087342	avgZ3	0.157777
		stddevY4	0.155477

Figura 12. Puntos del cuerpo - correlación extraversión grupal e individual

4.2.1 Categorización de variables objetivo

Por otra parte, se realiza un enfoque distinto al mencionado con anterioridad que se basa en clasificar a cada individuo en una categoría para posteriormente emplear la secuencia de movimiento de cada uno y los algoritmos de aprendizaje automático para intentar predecir dicha categoría. Para ello se han utilizado dos estrategias. Primero se han categorizado a las variables objetivo de forma individual usando umbrales. Por otro lado, se ha intentado usar un enforque más global para agrupar todas las variables de personalidad en superclases mediante clustering.

En la Tabla 4 se muestran los umbrales empleados para asignar la nueva etiqueta atendiendo a los valores que se obtengan en los datos.

Umbral	Nivel de Extraversión	Nivel de Neuroticismo
Bajo	<= 16	<= 16
Medio	< 16 y <= 21	< 16 y <= 21
Alto	>21	> 21

Tabla 4. Umbrales empleados para realizar la clasificación

En la Figura 13 se muestra el histograma para la variable neuroticismo individual empleando los umbrales descritos en la Tabla 4.

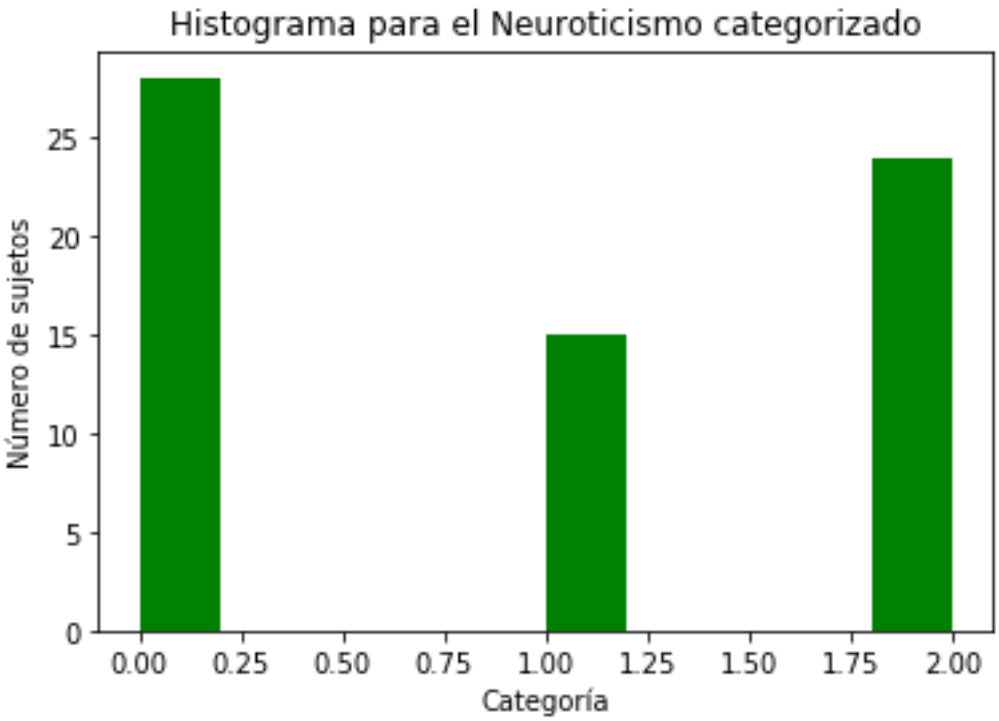


Figura 13. Histograma para el neuroticismo individual categorizado

Por otra parte, también se ha realizado un clustering de las variables objetivo neuroticismo, extraversión, apertura, cordialidad y responsabilidad grupales, con el fin de obtener una superclase que englobe varias de estas variables objetivo y, por tanto, permita agrupar los tipos de personalidad. Para este análisis se han empleado dos algoritmos, K-means y el clustering jerárquico. En la Figura 14 se muestra el mejor resultado obtenido para el algoritmo K-Means, el resto de las pruebas se muestran en el Anexo G. además, en dicha figura se muestra el método Average Silhouette, el cual permite medir la calidad del clustering, ya que este método determina que tan bien está cada elemento dentro de su cluster (mayor puntuación). Tal y como se puede observar, el número de clusters que mejor agrupa a las variables objetivo es 3.

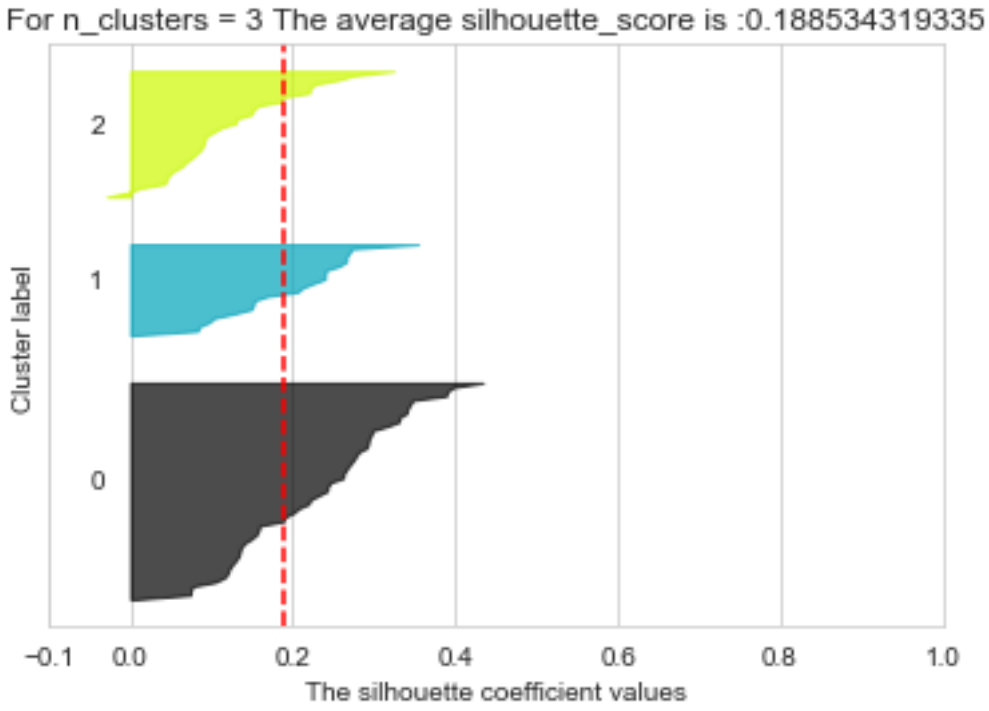


Figura 14. K-Means para 3 clusters

En la Tabla 5 se muestran los centroides obtenidos para estos clusters. Además, los rasgos de personalidad se han agrupado en tres superclases donde la primera tiene valores bajos de neuroticismo y medio-altos del resto de clases. En la segunda clase se puede observar que son todo valores medios. En la tercera clase se puede observar son valores medios excepto extraversión y apertura, los cuales son valores medio-altos.

Clúster	Neuroticismo grupal	Extraversión grupal	Apertura grupal	Cordialidad grupal	Responsabilidad grupal
1	16	34	33	32	36
2	27	25	24	27	28
3	24	36	34	24	26

Tabla 5. Centroides para K-means con 3 clusters

En la Figura 15 se muestra el mejor resultado obtenido para el clustering jerárquico, ya que para este clustering se han realizado diversas pruebas con diferentes métodos para unir los clusters, Anexo G, siendo la mejor la proporcionada por el método ward. Como se puede observar, el algoritmo empieza a juntar elementos atendiendo a dicho método, y finalmente se obtienen 4 grupos.

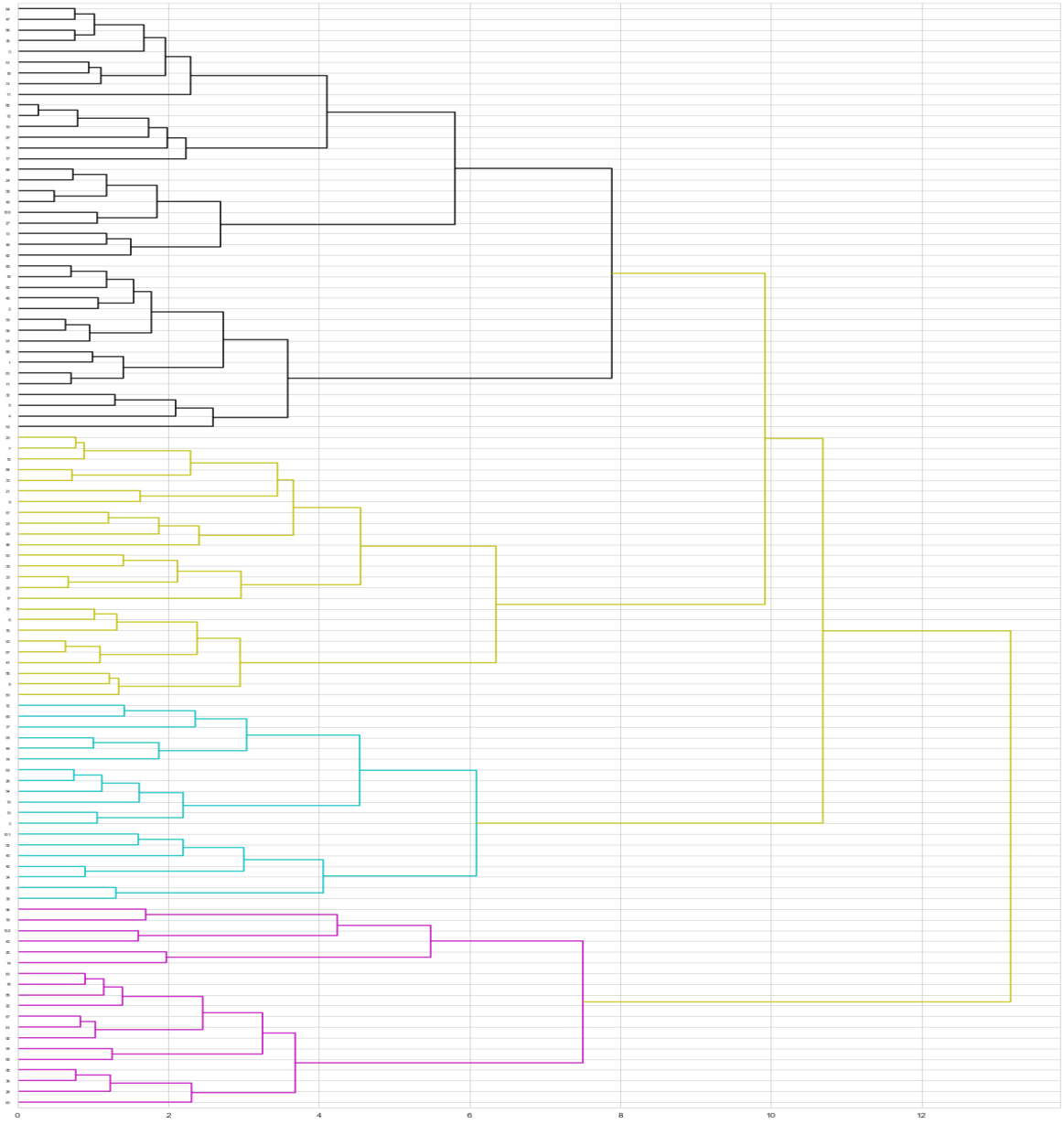


Figura 15. Clustering jerárquico con el método Ward

En la Tabla 6 se muestran los centroides obtenidos para estos clusters. Además, los rasgos de personalidad se han agrupado en cuatro superclases donde la primera tiene valores medios. En la segunda clase se puede observar que para extraversión, cordialidad y responsabilidad son valores medio-altos, mientras que para el resto son valores medios. En la tercera clase se puede observar son valores medios excepto extraversión y apertura, los cuales son valores medio-altos. En la cuarta clase se puede observar cómo son valores medio-altos excepto para neuroticismo que son valores bajos. Además, se puede comprobar como existen coincidencias entra el agrupamiento anterior y este. Existe una clase cuyos valores de neuroticismo son bajos, y el resto de las facetas de la personalidad tienen valores medio-altos. También, existen en ambos agrupamientos clases con valores medios para todas las facetas. Por último, existe una clase que destaca por tener valores medio-altos en extraversión y apertura, mientras que para el resto tiene valores medios.

Clúster	Neuroticismo grupal	Extraversión grupal	Apertura grupal	Cordialidad grupal	Responsabilidad grupal
1	28	25	25	24	24
2	19	36	20	31	32
3	24	35	38	28	26
4	16	33	35	30	37

Tabla 6. Centroides para clustering jerárquico

4.3 Resultados

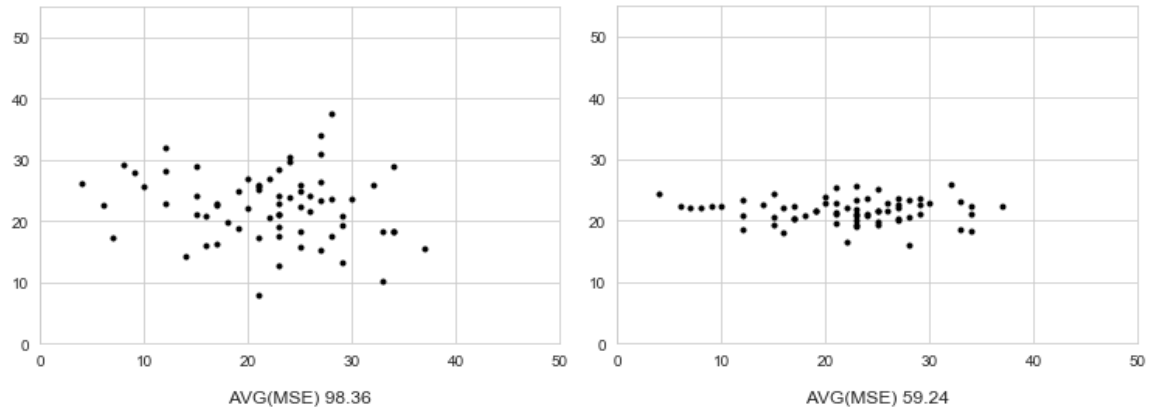
Las figuras que se muestran a partir de esta sección, es relevante mencionar que cada figura muestra en el eje X el valor del modelo real y en el eje Y, el valor predicho del conjunto de datos. Ambos ejes muestran los valores en las unidades de la variable objetivo. Cada punto representa un dato de test y los resultados de la validación cruzada se han juntado en una sola gráfica.

4.3.1 Características simples

En esta sección se muestran solo los resultados de las variables neuroticismo y extraversión ya que son los más representativos. El resto de los resultados se muestran en los anexos B (apertura), C (cordialidad), D (responsabilidad).

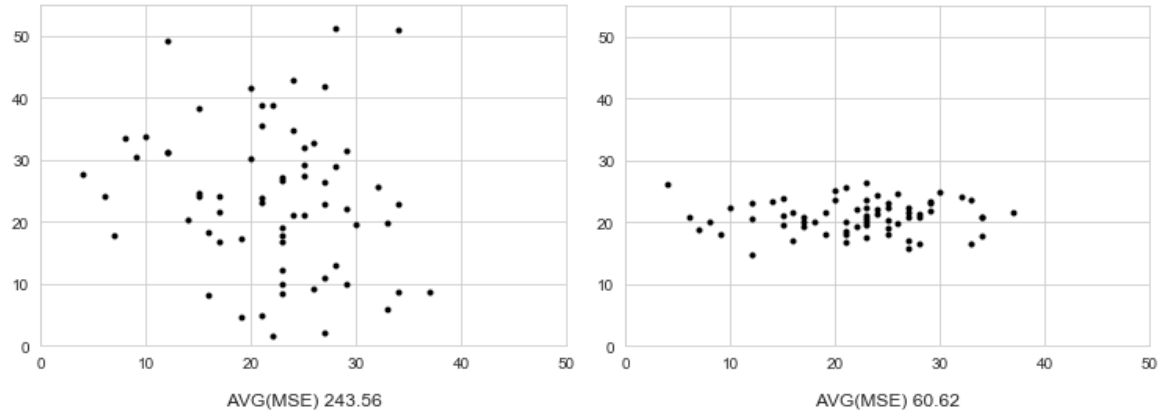
En la Figura 16 se muestran los resultados obtenidos para las características simples. Como se puede observar, el algoritmo Random Forest proporciona un mejor resultado ya que la media del MSE es menor, aunque aun así el error es alto. Sin embargo, este algoritmo predice un valor constante prácticamente para todos los puntos por lo que no es muy útil este resultado. En cuanto al neuroticismo, se puede observar que se obtiene un mejor resultado empleando todas las características simples en vez de solo las 6 más correlacionadas. En la extraversión se puede comprobar que no existen una gran diferencia entre emplear todas las características o solo las 6 más correlacionadas. Además, el neuroticismo grupal obtiene mejores resultados que el neuroticismo individual mientras que, en la extraversión ocurre lo contrario.

Regresión Lineal (a) y Random Forest Regression (b) para neuroticismo grupal.



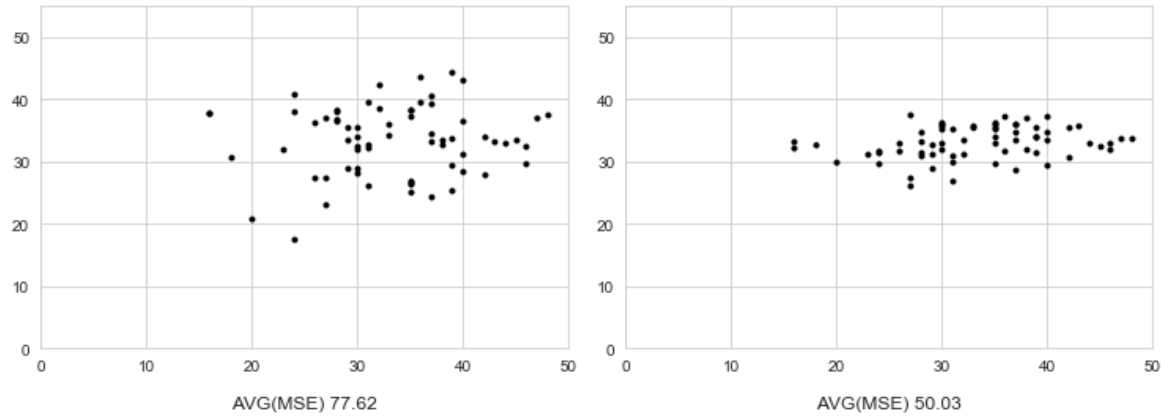
(a) (b)

Regresión Lineal (c) y Random Forest Regression (d) con los 6 puntos más correlacionados para neuroticismo grupal.



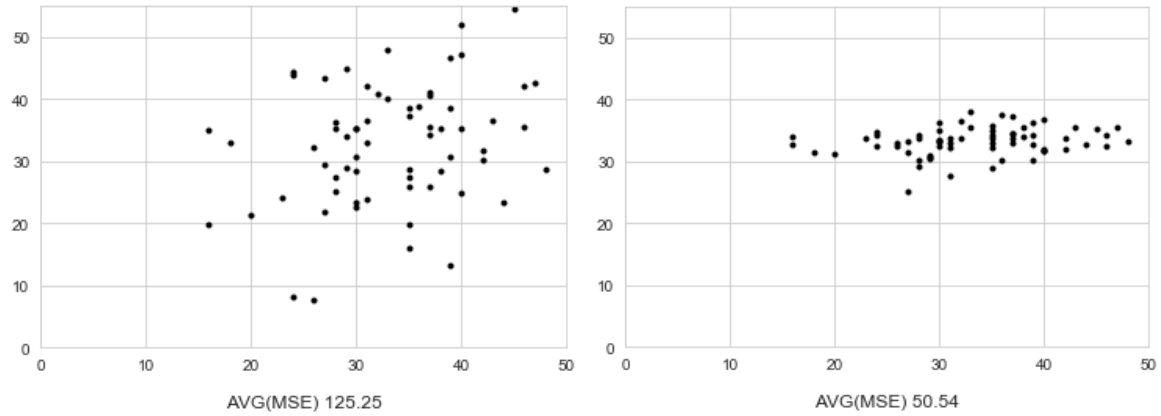
(c) (d)

Regresión Lineal (e) y Random Forest Regression(f) para extraversión individual.



(e) (f)

Regresión Lineal (g) y Random Forest Regression (h) con 6 variables más correlacionadas para extraversión individual



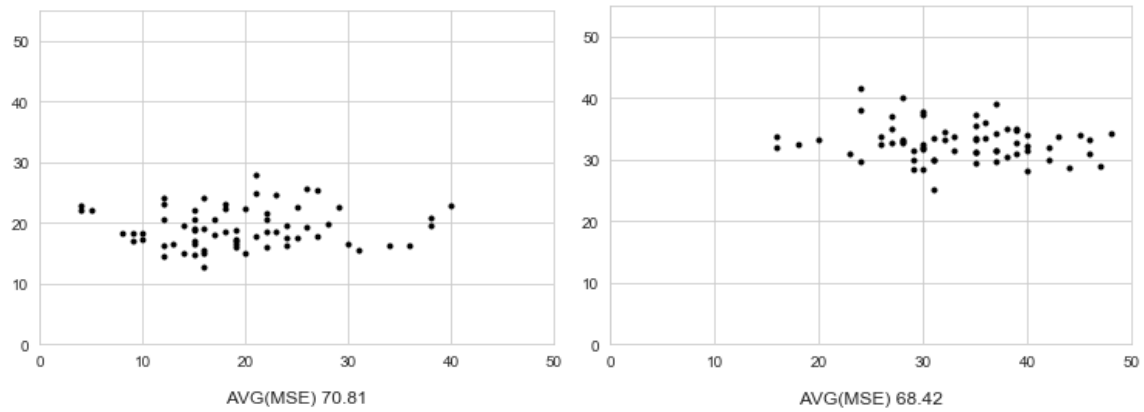
(g) (h)

Figura 16. Resultados de las características simples

4.3.2 Características compuestas

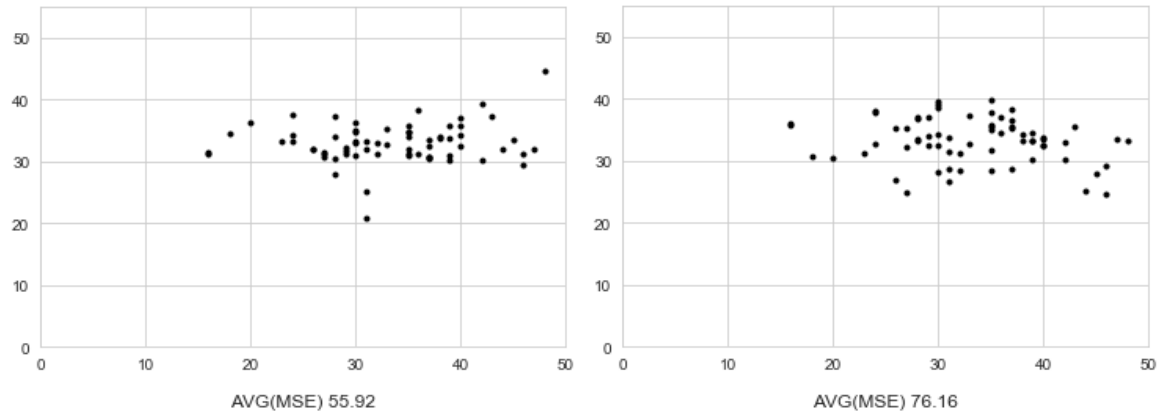
En la Figura 17 se muestran los resultados obtenidos para las características compuestas. Como se puede observar, el algoritmo Random Forest proporciona un mejor resultado ya que la media del MSE es menor, aunque aun así el error obtenido es alto. Como se puede observar, la distancia entre la cabeza y los pies es la que proporciona un mejor resultado, aunque el error es demasiado alto.

RandomForestRegression con distancia entre manos para neuroticismo individual (a) y extraversión individual (b).



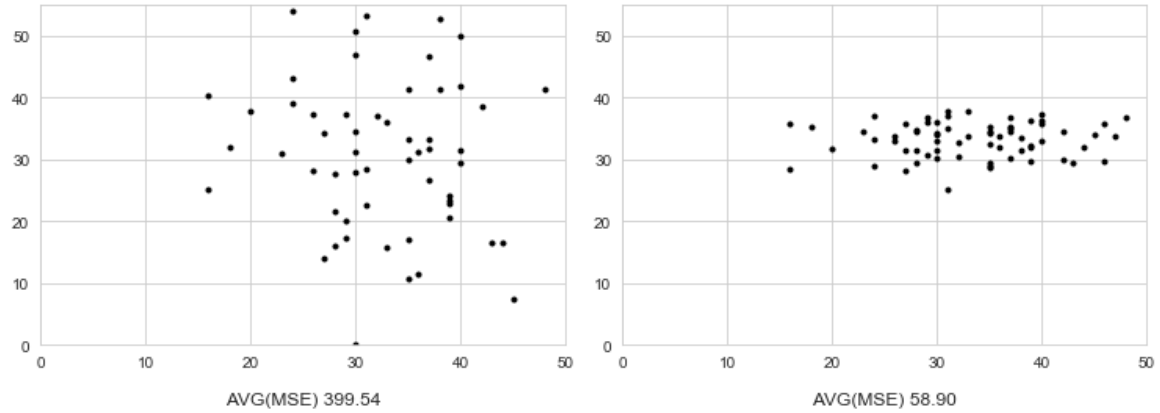
(a) (b)

Regresión Lineal con distancia entre cabeza y pies para extraversión individual (c) y RandomForestRegression con distancia euclídea entre manos para extraversión individual (d).



(c) (d)

Linear Regression (e) y Random Forest Regression (f) con intervalos con máxima y mínima movilidad para extraversión individual



(e) (f)

Figura 17. Resultados características complejas

4.3.3 Clustering

En esta sección se muestran los resultados de las pruebas del clustering sobre las distintas características y los clusters mencionados en la sección 4.2.

Umbrales

En la Figura 18 se muestra el resultado obtenido para la variable neuroticismo individual, empleando la distancia entre las manos, y el algoritmo Random Forest Classifier. Esta categorización de la variable neuroticismo es la mostrada en la Tabla 4. Para ello, se muestra la matriz de confusión, la cual tiene en el eje Y el número de elementos que se correspondían a cada una de las posibles categorías en los datos, y el eje X el número de elementos que se predicen para cada una de las posibles categorías. Por otra parte, se muestran los resultados en celdas cuyo color se corresponde con el número de aciertos, por lo que, cuanto más oscuro es el color mayor número de aciertos se han dado. Además, se puede observar que el

algoritmo no es demasiado preciso para neuroticismo medio, mientras que para neuroticismo bajo si detecta una mayoría de los individuos. Por lo tanto, tenemos un error de 1,37.

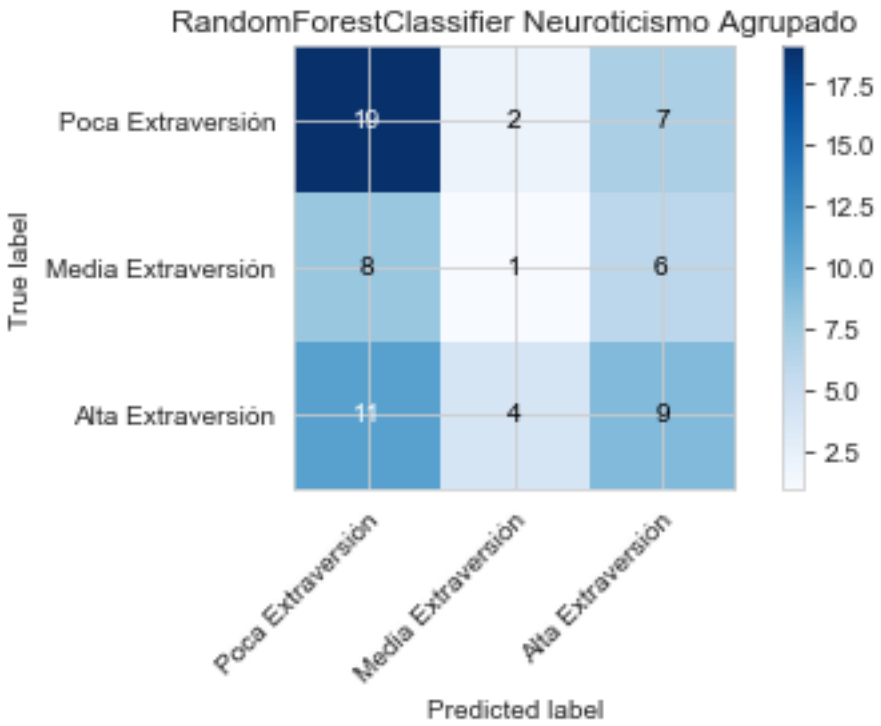


Figura 18. Random Forest Classifier – Neuroticismo

K-Means

En esta sección se muestra el mejor resultado obtenido en el clustering empleando las características simples y complejas mencionadas en la sección 3. El resto de los resultados se muestran el anexo G.

En la Figura 19 se muestra la matriz de confusión para el algoritmo regresión logística junto con las clases obtenidas mediante el clustering empleando K-Means con 3 clusters y como características se emplea la distancia euclídea entre las manos. Como se puede observar para las clases 1 y 3 se obtienen resultados precisos mientras que para la clase 2 no.

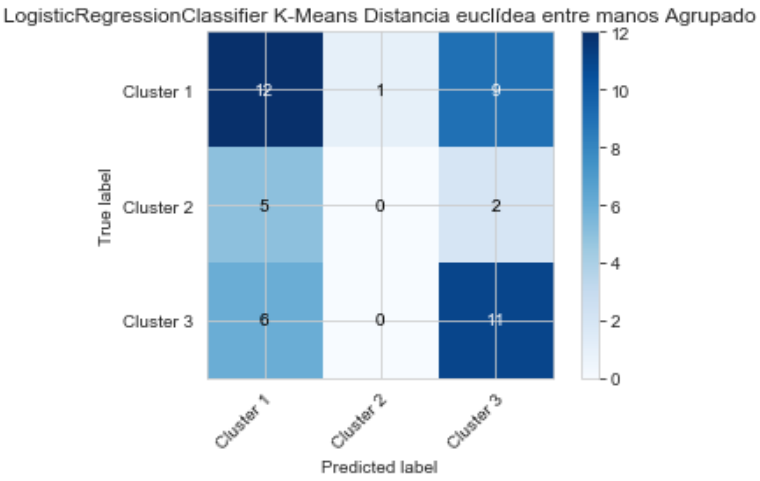


Figura 19. K-Means - Regresión logística con distancia euclídea entre manos

Clustering jerárquico

En esta sección se muestra el mejor resultado obtenido en el clustering empleando las características simples y complejas mencionadas en la sección 3. El resto de los resultados se muestran el anexo G.

En la Figura 20 se muestra la matriz de confusión para el algoritmo RandomForestClassifier, junto con las clases obtenidas mediante el clustering jerárquico que nos dio como resultado 4 clusters y como características se emplean las características simples. Como podemos observar, la clase 4 es la única en la que el algoritmo tiene cierta precisión mientras que en el resto de las clases no se obtiene precisión alguna. Por lo tanto, el error es de 2.66.

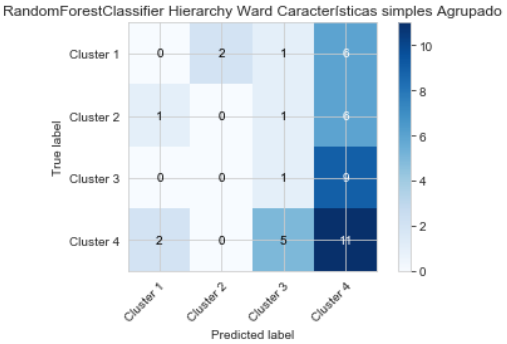


Figura 20. Clustering jerárquico - RandomForestClassifier con características simples

4.4 Discusión de resultados

En vista de los resultados expuestos anteriormente, se puede apreciar que en general el algoritmo Random Forest Regressor proporciona mejores resultados que el algoritmo de regresión lineal, tal y como se puede observar en la media del error cuadrático medio. Sin

embargo, los resultados proporcionados por este algoritmo no son demasiado buenos ya que se predice valores prácticamente constantes.

Por otra parte, también se puede apreciar que las características simples mejoran las aproximaciones con respecto a las características complejas. Esto es evidente, ya que en general, las características complejas empleadas (distancia entre las manos, distancia euclídea entre las manos, inclinación de la cabeza con respecto a los pies), ofrecen peores resultados, ya que su MSE es mayor. Además, esto puede estar causado por la pérdida excesiva de información que se produce en las características complejas, ya que ellas solo se centran ciertas partes del cuerpo y, por tanto, el resto de información que las demás aportaban se pierde, mientras que en las características simples esto no sucede ya que en ellas se emplean todas las partes de cuerpo.

También se puede apreciar que, entre todas las variables objetivo, la extraversión individual es la que muestra mejores resultados. Esto es evidente ya que el MSE en esta variable tiene a ser menor que en las demás variables, por lo que nos indica que esta variable tiende a tener un mayor impacto en los sujetos de la entrevista.

En cuanto a los intervalos de tiempo, se puede observar que no aportan ninguna diferencia con respecto a las demás pruebas y, que o bien no se han obtenido de forma correcta, es decir, los intervalos no están sincronizados adecuadamente, o bien durante esos intervalos de tiempo, el movimiento de los sujetos no es relevante en relación con los demás intervalos no analizados.

En cuanto al clustering es conveniente destacar la escasez de datos. En cuanto a la variable neuroticismo, se puede observar cómo existen una mayor diversidad de elementos para cada clase, figura 14, y, por tanto, esta variable nos ofrece una mejor visión de cómo afecta la distancia entre las manos a la hora de categorizar a los sujetos de las entrevistas. Por lo que, tal y como se puede observar en los resultados, esta característica no es decisiva a la hora de categorizar a los sujetos de las entrevistas según el neuroticismo de estos.

Por otro lado, se puede observar que con los clusters extraídos a partir de las variables objetivo y para el algoritmo de clustering K-means con 3 clusters, se puede observar cómo en general para las distintas características probadas los clusters 1 y 2, obtienen un buen porcentaje de acierto, destacando la distancia euclídea entre las manos, mientras que para el clúster 2 no se obtienen ningún buen resultado. En cuanto al algoritmo de clustering jerárquico, se obtienen 4 clusters y se puede observar como para cualquiera de las características empleadas existe una tendencia a clasificar todo al clúster número 4, dando como resultado una tasa de error muy alta.

5 Conclusiones y trabajo futuro

5.1 Conclusiones

El objetivo de este trabajo ha sido el de identificar patrones de movimiento que permitan predecir los distintos rasgos de personalidad de un individuo. Para ello se han analizado datos provenientes de entrevistas realizadas en el marco de un estudio de personalidad llevado a cabo por el Grupo de Investigación en Piscología y Ciencias del Deporte. Los datos que se han analizado incluyen las secuencias de movimiento de las distintas partes del cuerpo de los individuos durante las entrevistas, así como sus autoevaluaciones de personalidad términos de extraversión, neuroticismo, apertura, cordialidad y responsabilidad.

En primer lugar, se realizó un preprocesamiento de los datos para eliminar datos incompletos y uniformizar los datos. Posteriormente, una vez limpios los datos, se han extraído características a partir de las secuencias de movimiento de cada parte del cuerpo. Se han extraído dos tipos de atributos: simples y complejos. Los atributos simples: máximo, mínimo, media, desviación típica, movilidad y complejidad. Los complejos: distancia entre manos, distancia euclídea entre manos, inclinación de la cabeza con respecto a los pies. Finalmente se han aplicado técnicas de aprendizaje con el objetivo de encontrar las características más representativas para cada una de las variables objetivo.

Por otra parte, los resultados dejan claro que las características simples no son significativas a la hora de realizar la predicción de las facetas de la personalidad. Dado que el error obtenido es muy alto y que los algoritmos realizan predicciones constantes, no se puede obtener ninguna conclusión. En cuanto a las características complejas, los resultados de las distintas pruebas muestran que estas características no son relevantes y, por tanto, habría que definir nuevas características complejas enfocándose en la parte superior del cuerpo.

En cuanto al objetivo de clasificar a las personas según su grado de personalidad, los resultados no son concluyentes ya que para ciertas clases se obtienen buenos resultados, pero para el resto no. Para el objetivo de realizar un clustering agrupando todas las variables objetivo, se ha podido observar que los atributos empleados no son relevantes dado que siempre hay uno o varios clusters que las diferentes características empleadas no son capaces de aportar ninguna información relevante que ayude a los algoritmos a categorizarlos dentro de ese o esos clusters.

5.2 Trabajo futuro

Existen diversas vías que pueden seguirse para trabajos futuros. Una de ellas, es el empleo de algoritmos de análisis de secuencias temporales como puede ser Long Short Term Memory (LSTM), con el objetivo de analizar toda la información contenida en las secuencias de movimiento y que se pierde al utilizar estadísticos simples.

Otra vía para un trabajo futuro sería realizar un análisis espectral en el dominio de frecuencia, ya que estos resultados podrían ser empleados como entrada de un modelo basado en redes neuronales, las cuales han demostrado proporcionar mejores resultados que otros algoritmos de aprendizaje automático en los últimos tiempos.

Por otra parte, se podría intentar obtener nuevas características complejas mas representativas de la personalidad. Se seguiría en la línea del movimiento de las manos, pero se intentaría detectar patrones. Por ejemplo, si el sujeto se toca la cabeza cuando habla o si hace muchos aspavientos, etc.

Bibliografía

[1] Wikipedia. Psicología. https://es.wikipedia.org/wiki/Psicolog%C3%ADa. Último acceso 08 de junio de 2019.

[2] Roth, P. L., Bobko, P., Van Iddekinge, C. H., & Thatcher, J. B. (2016). Social media in employee-selection-related decisions: A research agenda for uncharted territory. Journal of Management,42(1), 269-298.

[3] Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. Criminal Justice and Behavior, 36(1), 21-40.

[4] Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems,2(1), 3 doi: 10.1186/2047-2501-2-3.

[5] North, M. (1972). Personality Assessment through movement. Boston: Plays, Inc.

[6] Andrea Kleinsmith and Nadia Bianchi-Berthouze (2013). Affective Body Expression Perception and Recognition: A Survey.

[7] FEBI assessment. Eysenck's Personality Inventory (EPI) (Extroversion/Introversion). http://febiassessment.com/test/eysencks-personality-inventory-epi-extroversionintroversion/. Último acceso 08 de junio de 2019.

[8] Molina, Xavier. Neurosis (neuroticismo): causas, síntomas y características. https://psicologiaymente.com/clinica/neurosis-neuroticismo-causas-sintomas. Último acceso 08 de junio de 2019.

[9] Psychologist World. Extraversion and Introversion. https://www.psychologistworld.com/in fluence-personality/extraversion-introversion. Último acceso 08 de junio de 2019.

[10] Barry Kaufman, Scott.The Difference between ExtrAversion and ExtrOversion. https://blogs.scientificamerican.com/beautiful-minds/the-difference-between-extraversion-and-extroversion/?redirect=1. Último acceso 08 de junio de 2019.

[11] Miravalles, Javier. La responsabilidad. http://www.javiermiravalles.es/EV/La%20Responsabilidad.html. Último acceso 08 de junio de 2019.

[12] Giraldo, Katherine. Cordialidad — Valor Personal. https://www.kathegiraldo.com/cordialidad/. Último acceso 08 de junio de 2019.

[13] Muñoz, Ana. Los 5 rasgos de la personalidad: Apertura a la experiencia. https://www.aboutespanol.com/los-5-rasgos-de-la-personalidad-apertura-a-la-experiencia-2396167. Último acceso 08 de junio de 2019.

[14] Jowitt, Tom. Tales In Tech History: Microsoft Kinect. https://www.silicon.co.uk/e-innovation/microsoft-kinect-history-226781. Último acceso 08 de junio de 2019.

[15] Corden, Jez. Farewell, dear sweet Kinect. https://www.windowscentral.com/ode-kinect. Último acceso 08 de junio de 2019.

[16] Cong, Robert & Winters, Ryan. How Does The Xbox Kinect Work. https://www.jameco.com/jameco/workshop/howitworks/xboxkinect.html. Último acceso 08 de junio de 2019.

[17] Warren, Tom. A closer look at Microsoft's new Kinect sensor. https://www.theverge.com/2019/2/25/18239860/microsoft-kinect-azure-dk-hands-on-mwc-2019. Último acceso 08 de junio de 2019.
[18] Microsoft. Azure Kinect DK. https://azure.microsoft.com/en-us/services/kinect-dk/. Último acceso 08 de junio de 2019.
[19] Amazon. Computación a gran escala y conjuntos de datos grandes. https://media.amazonwebservices.com/es/DataSheet_Architecture/RefArch_LargeScale_5Ar.pdf. Último acceso 08 de junio de 2019.
[20] Microsoft Azure. Azure Kubernetes Service (AKS). https://azure.microsoft.com/en-us/services/kubernetes-service/. Último acceso 08 de junio de 2019.
[21] Databricks. Databricks Unified Analytics Platform. https://databricks.com/product/unified-analytics-platform. Último acceso 08 de junio de 2019.
[22] Apache Hadoop. Apache Hadoop. https://hadoop.apache.org/. Último acceso 08 de junio de 2019.
[23] Apuntes de clase. Computación a gran escala. Máster Universitario en Ingeniería Informática.
[24] Spark – Apache. Spark SQL, DataFrames and Datasets Guide. https://spark.apache.org/docs/latest/sql-programming-guide.html. Último acceso 08 de junio de 2019.
[25] R. Dillmann, O. Rogalla, M. Ehrenmann, R. Zöliner y M. Bordegoni. Learning Robot Behaviour and Skills Based on Human Demonstration and Advice: The Machine Learning Paradigm.
[26] Valerie Guralnik y Karen Zita Haigh. Learning Models of Human Behaviour with Sequential Patterns.
[27] Maja Pantic. Machine analysis of facial behaviour: naturalistic and dynamic behaviour.
[28] Benson Edwin Raj y Annie Portia. Analysis on credit card fraud detection methods.
[29] Edward Rosten y Tom Drummond. Machine Learning for High-Speed Corner Detection.
[30] Michael J. Pazzani y Daniel Billsus. Content-Based Recommendation Systems.
[31] Apuntes de clase. Sistemas basados en conocimiento y minería de datos. Máster Universitario en Ingeniería Informática. Último acceso 08 de junio de 2019.
[32] Recuero, Paloma. Los 2 tipos de aprendizaje en Machine Learning: supervisado y no supervisado. https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/. Último acceso 08 de junio de 2019.
[33] Anilu Franco-Arcega1, Jesús Ariel Carrasco-Ochoa, Guillermo Sánchez-Díaz y José Francisco Martínez-Trinidad. Decision Tree based Classifiers for Large Datasets
[34] Statistics How To. Linear Regression: Simple Steps and Video — Find the Equation, Coefficient and Slope. https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/. Último acceso 08 de junio de 2019.

[35] Marín Diazaraque, Juan Miguel. Análisis de regresión lineal. http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/GuiaSPSS/18reglin.pdf. Último acceso 08 de junio de 2019.
[36] Wikipedia. Regresión logística. https://es.wikipedia.org/wiki/Regresi%C3%B3n_log%C3%ADstica. Último acceso 08 de junio de 2019.
[37] Breiman, Leo (2001). "Random Forests". Machine Learning 45 (1): 5-32. doi:10.1023/A:1010933404324.
[38] Piech, Chris. K Means. https://stanford.edu/~cpiech/cs221/handouts/kmeans.html. Último acceso 08 de junio de 2019.
[39] Scipy.org. Hierarchical clustering. https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html. Último acceso 08 de junio de 2019.
[40] Wasilewska, Anita. APRIORI Algorithm. https://www3.cs.stonybrook.edu/~cse634/lecture_notes/07apriori.pdf. Último acceso 08 de junio de 2019.
[41] Acosta, Jr y McCrae, Robert R. NEO FFI, Inventario Neo reducido de cinco factores.
[42] Benítez Aldás, Marcos Raphael. Estudio y análisis de métodos para la extracción de características y clasificación de emociones basados en EEG, 48-49.

Anexos

A Características simples y descripciones

En este anexo se muestran los identificadores empleados junto con sus correspondientes descripciones.

En la Tabla 7, se muestran las características simples junto con su descripción.

Identificador	Descripción
maxX1 – X25	Máximo valor del eje X para los puntos 1 a 25
minX1 – X25	Mínimo valor del eje X para los puntos 1 a 25
avgX1 – X25	Media del eje X para los puntos 1 a 25
stddevX1 – X25	Desviación estándar del eje X para los puntos 1 a 25
MX1 – X25	Movilidad del eje X para los puntos 1 a 25
CX1 – X25	Complejidad del eje X para los puntos 1 a 25
SumDiffAvgX1 – X25	Suma de las diferencias con respecto a la media para el eje X para los puntos 1 a 25
maxY1 – Y25	Máximo valor del eje Y para los puntos 1 a 25
minY1 – Y25	Mínimo valor del eje Y para los puntos 1 a 25
avgY1 – Y25	Media del eje Y para los puntos 1 a 25
stddevY1 – Y25	Desviación estándar del eje Y para los puntos 1 a 25
MY1 – Y25	Movilidad del eje Y para los puntos 1 a 25
CY1 – Y25	Complejidad del eje Y para los puntos 1 a 25
SumDiffAvgY1 – Y25	Suma de las diferencias con respecto a la media para el eje Y para los puntos 1 a 25
maxZ1 – Z25	Máximo valor del eje Z para los puntos 1 a 25
minZ1 – Z25	Mínimo valor del eje Z para los puntos 1 a 25
avgZ1 – Z25	Media del eje Z para los puntos 1 a 25
stddevZ1 – Z25	Desviación estándar del eje Z para los puntos 1 a 25
MZ1 – Z25	Movilidad del eje Z para los puntos 1 a 25
CZ1 – Z25	Complejidad del eje Z para los puntos 1 a 25
SumDiffAvgZ1 – Z25	Suma de las diferencias con respecto a la media para el eje Z para los puntos 1 a 25

Tabla 7. Características simples - descripción

B Estadísticas variable apertura

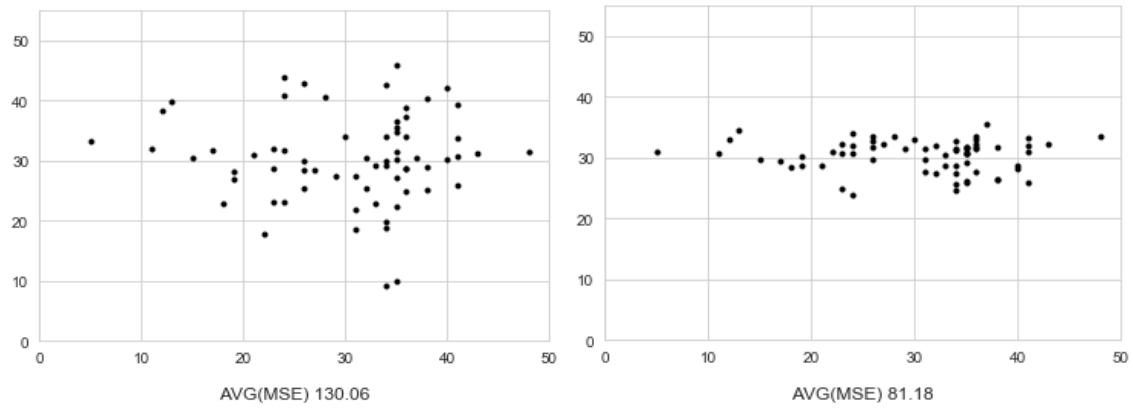
En este anexo se muestran las estadísticas relacionadas con la variable objetivo apertura.

En la Figura 21 se muestra la correlación para los puntos del cuerpo tanto para la variable objetivo apertura grupal como apertura individual, ordenados de forma descendente. Además, se puede observar que los puntos del cuerpo P3, P4, P5, P8, P9 y P12, son los que presentan una correlación más alta con respecto a las variables apertura individual y grupal. Esto tiene sentido ya que estos puntos se corresponden a la zona superior de cuerpo que es con la que la persona más gesticula, y dado que la apertura es una faceta de la personalidad que incluye el entusiasmo de una persona.

total_APER_grupal	1.000000	total_APER_indiv	1.000000
minY3	0.351555	CY5	0.244658
minY9	0.347204	SumDiffAvgY4	0.219921
minY4	0.345536	SumDiffAvgZ12	0.215023
minY5	0.324661	SumDiffAvgZ8	0.197719
CZ3	0.250985	SumDiffAvgZ9	0.196923
minZ12	0.205414	SumDiffAvgY5	0.190372
minY8	0.202799	SumDiffAvgZ4	0.183038
MX8	0.195666	CY4	0.176542
MX4	0.194470	SumDiffAvgY3	0.172485
MX9	0.193177	SumDiffAvgZ3	0.154553
minY12	0.189787	SumDiffAvgZ5	0.153395
minZ4	0.183404	SumDiffAvgY9	0.141065
minZ9	0.167980	SumDiffAvgX8	0.134806
minZ5	0.165290	SumDiffAvgX5	0.126656
MX3	0.164042	stddevZ8	0.125612
CY5	0.161789	MY8	0.124767
minZ3	0.147704	MX9	0.118125
minZ8	0.144459	MX8	0.117536
CY4	0.141945	stddevZ12	0.112366
MX12	0.123903	CX12	0.100877
CY12	0.121658	stddevZ9	0.100245
MX5	0.117498	minZ12	0.096951
maxX8	0.111301	stddevZ4	0.091935
CX3	0.102248	maxY8	0.090905
avgX12	0.095158	stddevX5	0.081784
		stddevX8	0.080896

Figura 21. Puntos del cuerpo - correlación apertura grupal e individual

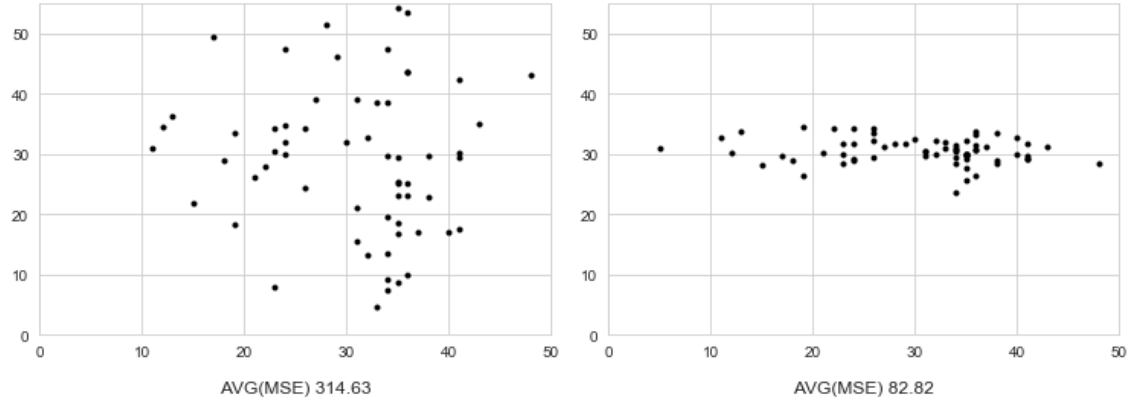
En la Figura 22 se muestra el resultado obtenido para la variable apertura grupal, empleando las características simples mostradas en el Anexo A y los algoritmos Linear Regression y Random Forest Regression.



(a) (b)

Figura 22. Regresión Lineal (a) y Random Forest Regression (b) para apertura grupal

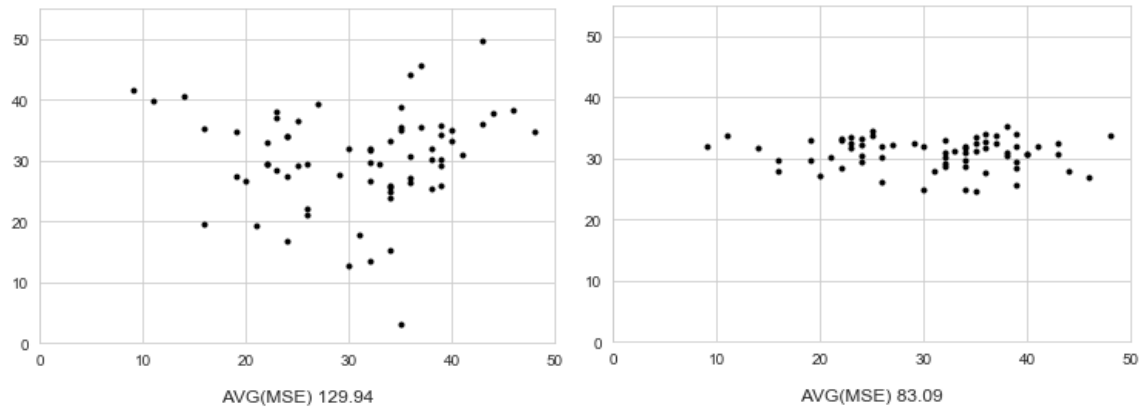
En la Figura 23 se muestra el resultado obtenido para la variable apertura grupal, empleando los seis puntos del cuerpo más correlacionados con dicha variable y los algoritmos Linear Regression y Random Forest Regressor.



(a) (b)

Figura 23. Regresión Lineal (a) y Random Forest Regressor (b) con los 6 puntos más correlacionados para apertura grupal

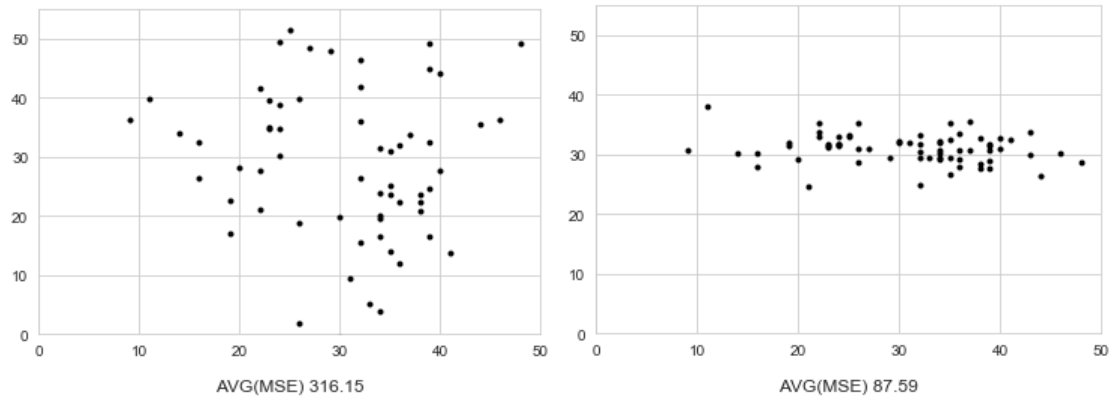
En la Figura 24 se muestran los resultados obtenidos para la variable apertura individual, empleando las características simples mostradas en el Anexo A y los algoritmos Linear Regression y Random Forest Regression.



(a) (b)

Figura 24. Regresión Lineal (a) y Random Forest Regression (b) para apertura individual

En la Figura 25 se muestra el resultado obtenido para la variable apertura individual, empleando los seis puntos del cuerpo más correlacionados con dicha variable y los algoritmos Linear Regression y Random Forest Regressor.



(a) (b)

Figura 25. Regresión Lineal (a) y Random Forest Regressor (b) con los 6 puntos más correlacionados para apertura individual

C Estadísticas variable cordialidad

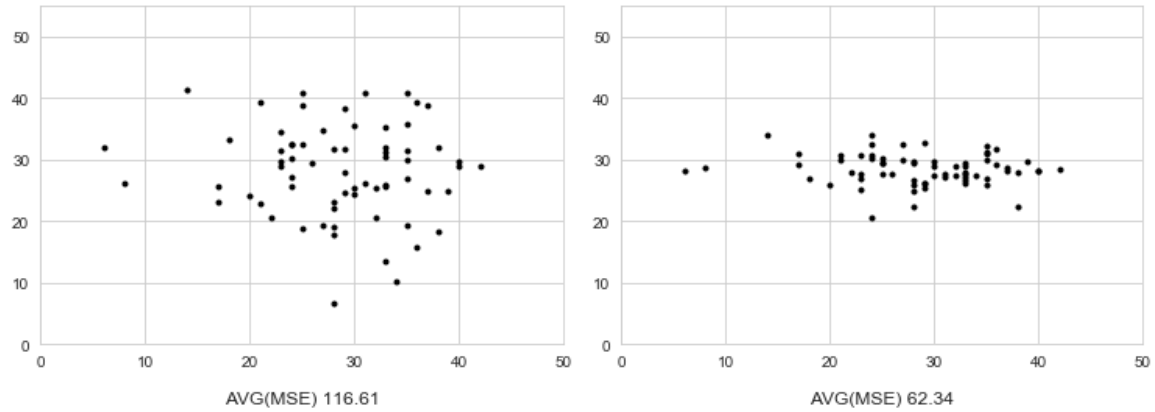
En este anexo se muestran las estadísticas relacionadas con la variable objetivo cordialidad.

En la Figura 26 se muestra la correlación para los puntos del cuerpo tanto para la variable objetivo cordialidad grupal como cordialidad individual, ordenados de forma descendente. Además, se puede observar que los puntos del cuerpo P3, P4, P5, P8, P9 y P12, son los que presentan una correlación más alta con respecto a las variables cordialidad individual y grupal.

total_CORD_grupal	1.000000	total_CORD_indiv	1.000000
CZ3	0.233043	maxZ8	0.230537
MX12	0.123135	maxZ5	0.215041
MZ8	0.119478	maxZ3	0.187011
maxZ5	0.116729	stddevY5	0.182341
avgY8	0.115186	stddevZ5	0.180763
maxZ8	0.109106	maxZ4	0.171056
maxY8	0.104397	CZ4	0.169024
avgY12	0.098421	SumDiffAvgY3	0.159660
MZ5	0.094991	stddevY3	0.156236
maxY12	0.091232	stddevZ3	0.151806
CZ4	0.090823	stddevY4	0.149162
minY9	0.078256	SumDiffAvgY9	0.147990
CX5	0.068073	stddevZ4	0.147773
CZ5	0.067479	stddevY9	0.138326
maxZ12	0.065658	SumDiffAvgZ5	0.137010
maxZ3	0.063895	SumDiffAvgY5	0.136496
maxZ4	0.062813	avgZ4	0.131626
CY5	0.057303	SumDiffAvgY4	0.123133
minY4	0.056457	stddevZ9	0.117658
minY3	0.055685	avgZ3	0.116488
MZ12	0.048353	CX4	0.111535
MX8	0.039273	MZ5	0.109175
avgX8	0.037693	stddevZ8	0.107745
MX4	0.033744	stddevZ12	0.106410
MZ9	0.032159	maxZ9	0.103183
CX4	0.030017	avgZ9	0.101353

Figura 26. Puntos del cuerpo - correlación cordialidad grupal e individual

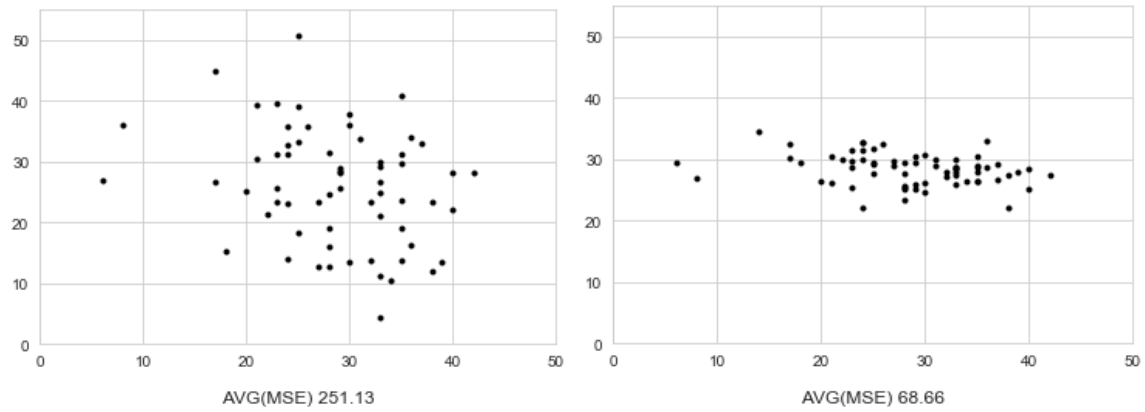
En la Figura 27 se muestra el resultado obtenido para la variable cordialidad grupal, empleando las características simples mostradas en el Anexo A y los algoritmos Linear Regression y Random Forest Regression.



(a) (b)

Figura 27. Regresión Lineal (a) y Random Forest Regression (b) para cordialidad grupal

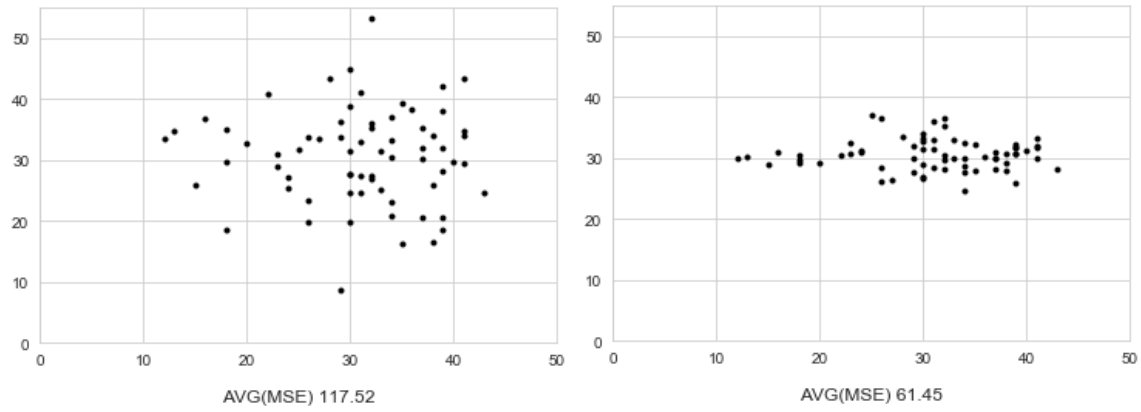
En la Figura 28 se muestra el resultado obtenido para la variable cordialidad grupal, empleando los seis puntos del cuerpo más correlacionados con dicha variable y los algoritmos Linear Regression y Random Forest Regression.



(a) (b)

Figura 28. Regresión Lineal (a) y Random Forest Regression (b) con los 6 puntos más correlacionados para cordialidad grupal

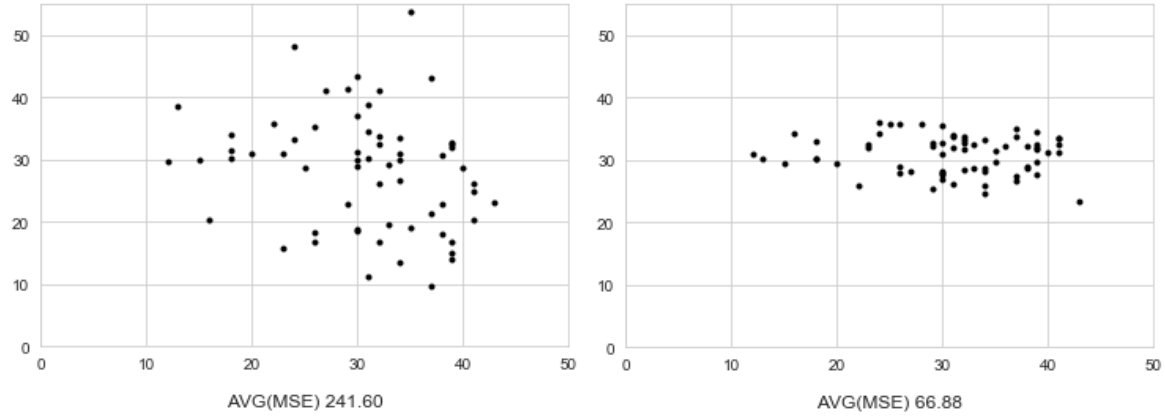
En la Figura 29 se muestran los resultados obtenidos para la variable cordialidad individual, empleando las características simples mostradas en el Anexo A y los algoritmos Linear Regression y Random Forest Regression.



(a) (b)

Figura 29. Regresión Lineal (a) y Random Forest Regression (b) para cordialidad individual

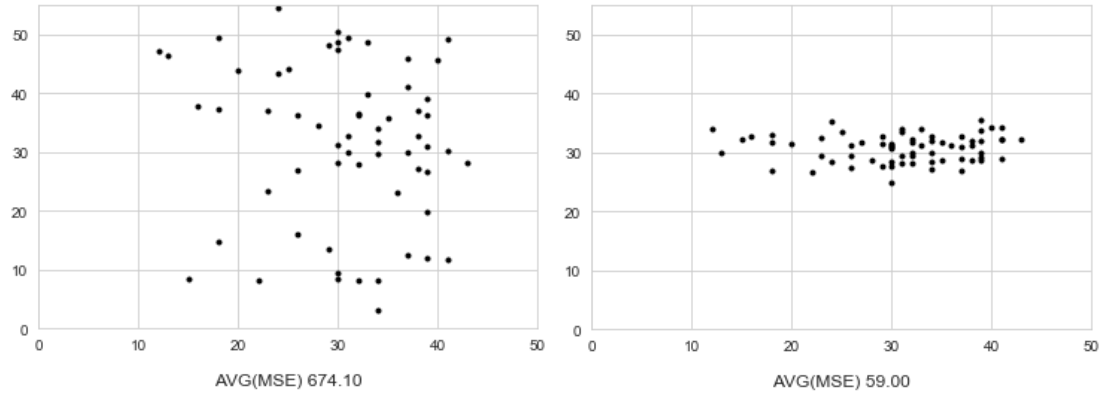
En la Figura 30 se muestra el resultado obtenido para la variable cordialidad individual, empleando los seis puntos del cuerpo más correlacionados con dicha variable y los algoritmos Linear Regression y Random Forest Regression.



(a) (b)

Figura 30. Regresión Lineal (a) y Random Forest Regression (b) con los 6 puntos más correlacionados para cordialidad individual

En la Figura 31 se muestra el resultado obtenido para la variable cordialidad individual, empleando los intervalos de tiempo donde la movilidad de la distancia entre los puntos 8 y 9, es máxima y mínima, y los algoritmos Linear Regression y Random Forest Regression.



(a) (b)

Figura 31. Linear Regression (a) y Random Forest Regression (b) con intervalos con máxima y mínima movilidad para cordialidad individual

D Estadísticas variable responsabilidad

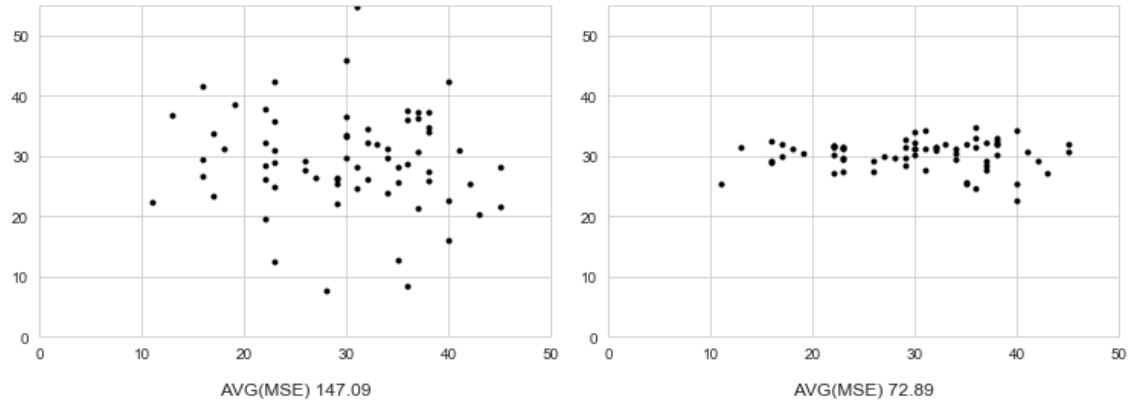
En este anexo se muestran las estadísticas relacionadas con la variable objetivo responsabilidad.

En la Figura 32 se muestra la correlación para los puntos del cuerpo tanto para la variable objetivo responsabilidad grupal como responsabilidad individual, ordenados de forma descendente. Además, se puede observar que los puntos del cuerpo P3, P4, P5, P8, P9 y P12, son los que presentan una correlación más alta con respecto a las variables responsabilidad individual y grupal.

total_RESP_grupal	1.000000	total_RESP_indiv	1.000000
CZ4	0.234370	CZ4	0.273307
MZ12	0.162178	maxZ8	0.182930
CZ12	0.137805	stddevY12	0.177038
minY3	0.136526	CZ12	0.138536
minY4	0.125516	MZ8	0.133678
CY9	0.125115	MZ12	0.131767
maxZ8	0.125094	maxZ5	0.121552
minY5	0.124497	MZ9	0.116906
avgX8	0.120976	maxZ12	0.101598
avgX4	0.118546	MZ3	0.098993
minY9	0.118330	MZ4	0.092906
CZ3	0.117574	maxZ9	0.090575
stddevY12	0.113808	MZ5	0.090543
avgX3	0.106149	maxZ3	0.083345
MZ9	0.103068	avgX8	0.081461
avgX9	0.092932	CY9	0.077224
avgX5	0.088090	maxZ4	0.075834
CZ9	0.084478	avgX4	0.071305
MZ8	0.082685	SumDiffAvgY12	0.070560
maxZ5	0.066500	minY3	0.069879
MX12	0.060752	avgZ4	0.065235
MZ3	0.059617	avgX3	0.060211
avgZ5	0.058320	avgZ3	0.057011
MZ4	0.057776	avgZ5	0.056931
		avgX9	0.053579

Figura 32. Puntos del cuerpo - correlación responsabilidad grupal e individual

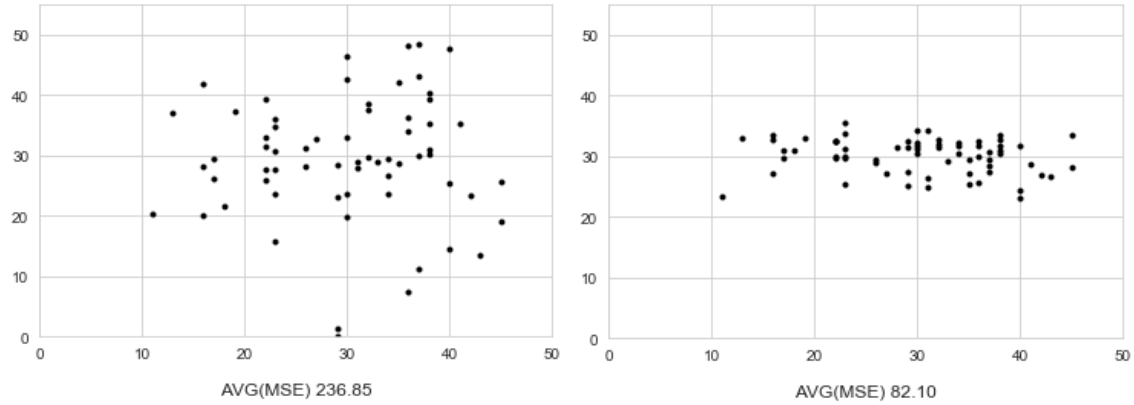
En la Figura 33 se muestran los resultados obtenidos para la variable responsabilidad grupal, empleando las características simples mostradas en el Anexo A y los algoritmos Linear Regression y Random Forest Regression.



(a) (b)

Figura 33. Regresión Lineal (a) y Random Forest Regression (b) para responsabilidad grupal

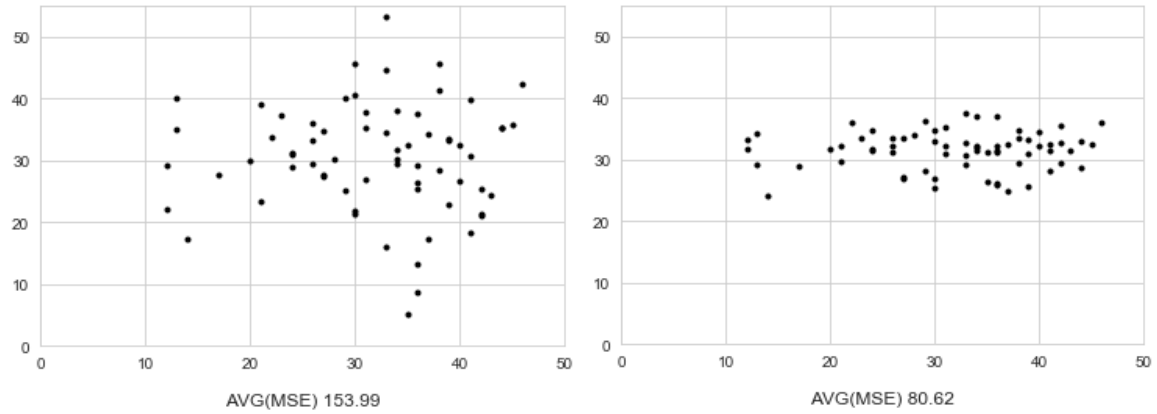
En la Figura 34 se muestra el resultado obtenido para la variable cordialidad grupal, empleando los seis puntos del cuerpo más correlacionados con dicha variable y los algoritmos Linear Regression y Random Forest Regression.



(a) (b)

Figura 34. Regresión Lineal (a) y Random Forest Regression (b) con los 6 puntos más correlacionados para responsabilidad grupal

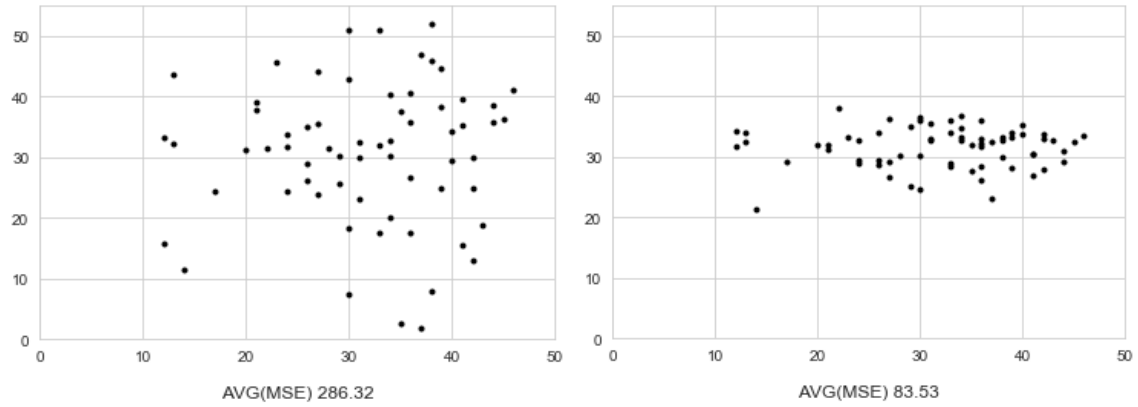
En la Figura 35 se muestran los resultados obtenidos para la variable responsabilidad individual, empleando las características simples mostradas en el Anexo A y los algoritmos Linear Regression y Random Forest Regression.



(a) (b)

Figura 35. Regresión Lineal (a) y Random Forest Regression (b) para responsabilidad individual

En la Figura 36 se muestra el resultado obtenido para la variable responsabilidad individual, empleando los seis puntos del cuerpo más correlacionados con dicha variable y los algoritmos Linear Regression y Random Forest Regression.



(a) (b)

Figura 36. Regresión Lineal (a) y Random Forest Regression (b) con los 6 puntos más correlacionados para responsabilidad individual

E Estadísticas variable neuroticismo

En este anexo se muestran las estadísticas relacionadas con la variable objetivo neuroticismo.

En la Figura 37, se muestra la matriz de correlación correspondiente a la variable neuroticismo individual, empleando como características la distancia entre las manos.

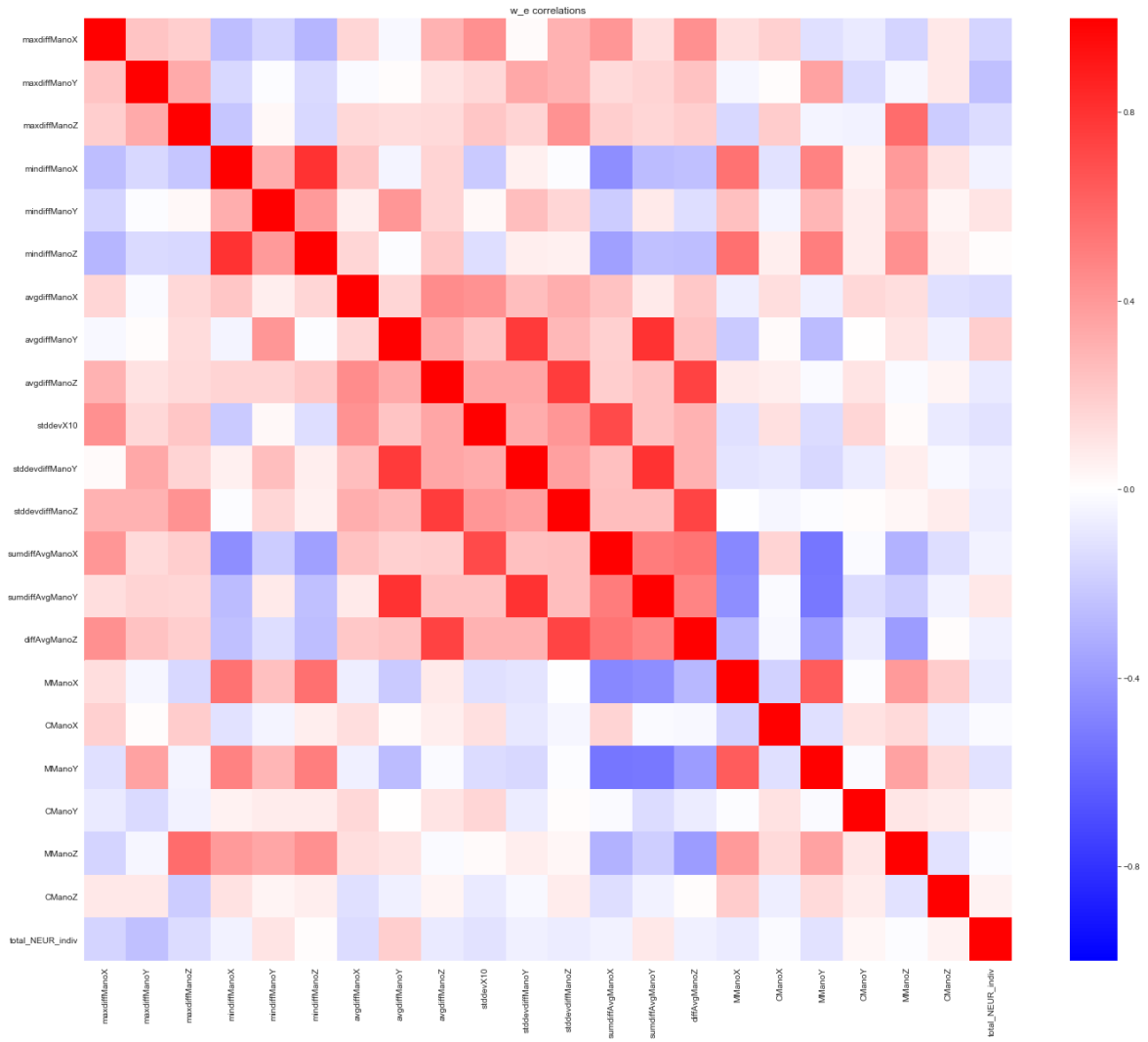


Figura 37. Matriz de correlación - Neuroticismo individual distancia entre manos

En la Figura 38, se muestra la matriz de correlación correspondiente a la variable neuroticismo individual categorizado, empleando la distancia entre las manos.

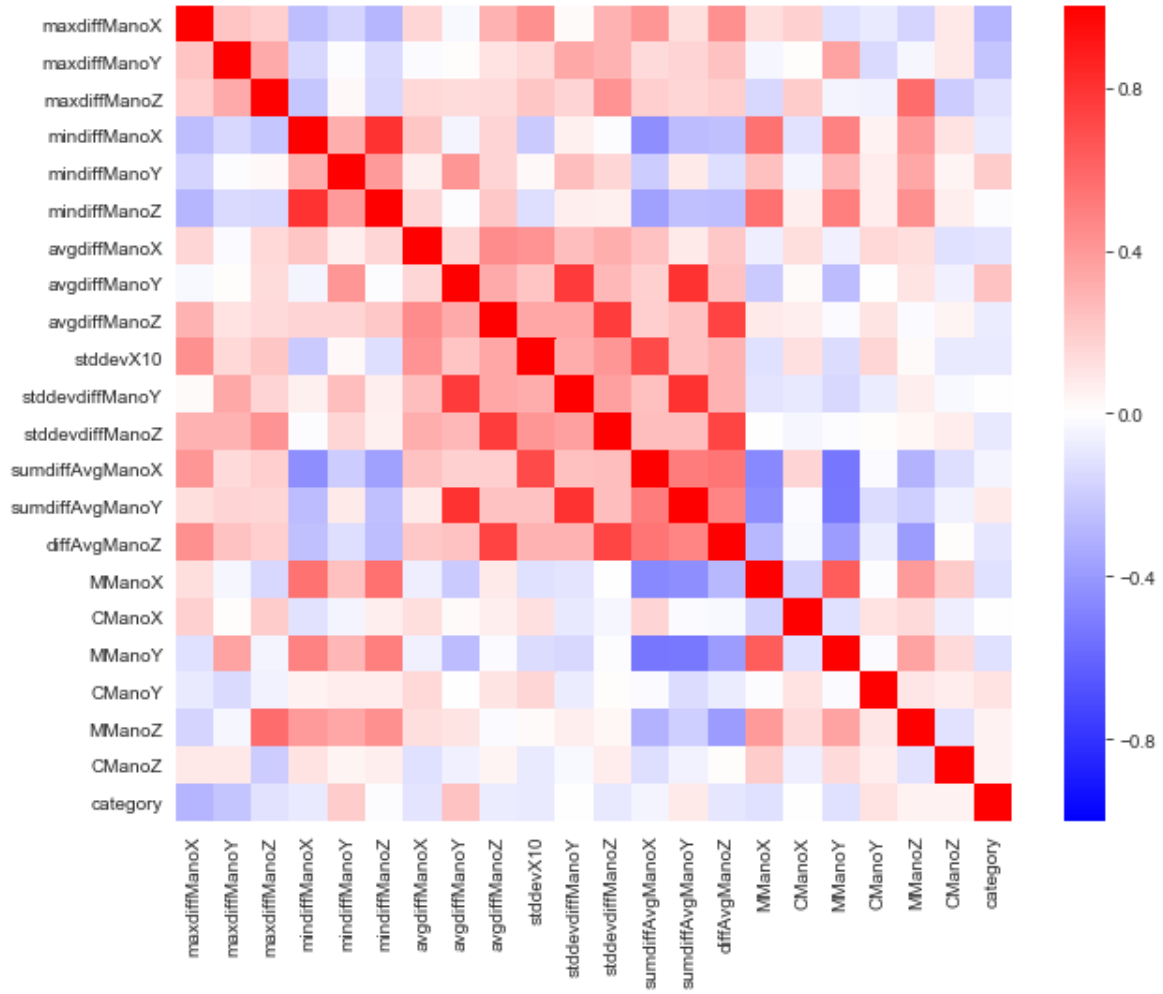


Figura 38. Matriz de correlación – Neuroticismo individual categorizado

F Estadísticas variable extraversión

En este anexo se muestran las estadísticas relacionadas con la variable objetivo extraversión.

En la Figura 39 se muestra la matriz de correlación correspondiente a la variable extraversión individual categorizada, empleando la distancia entre las manos.

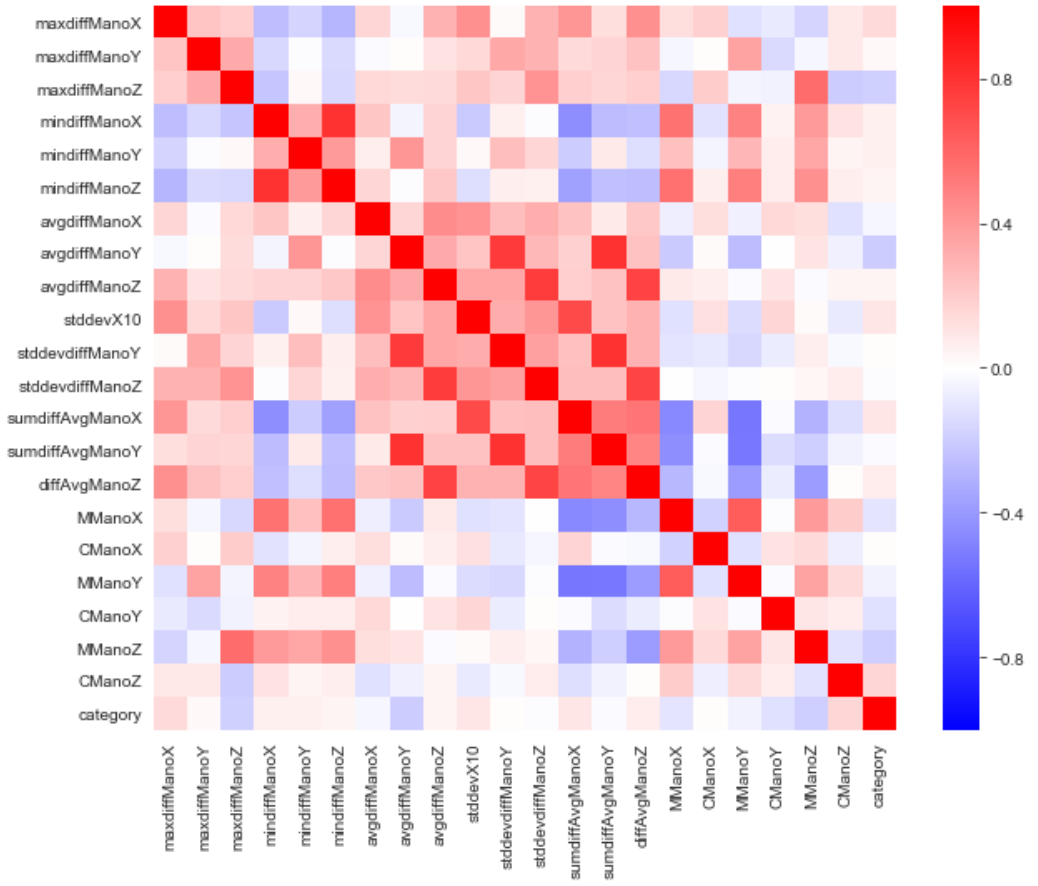


Figura 39. Matriz de correlación – Extraversión individual categorizada

En la Figura 40 se muestra la matriz de correlación correspondiente a la variable extraversión individual, empleando la distancia entre las manos.

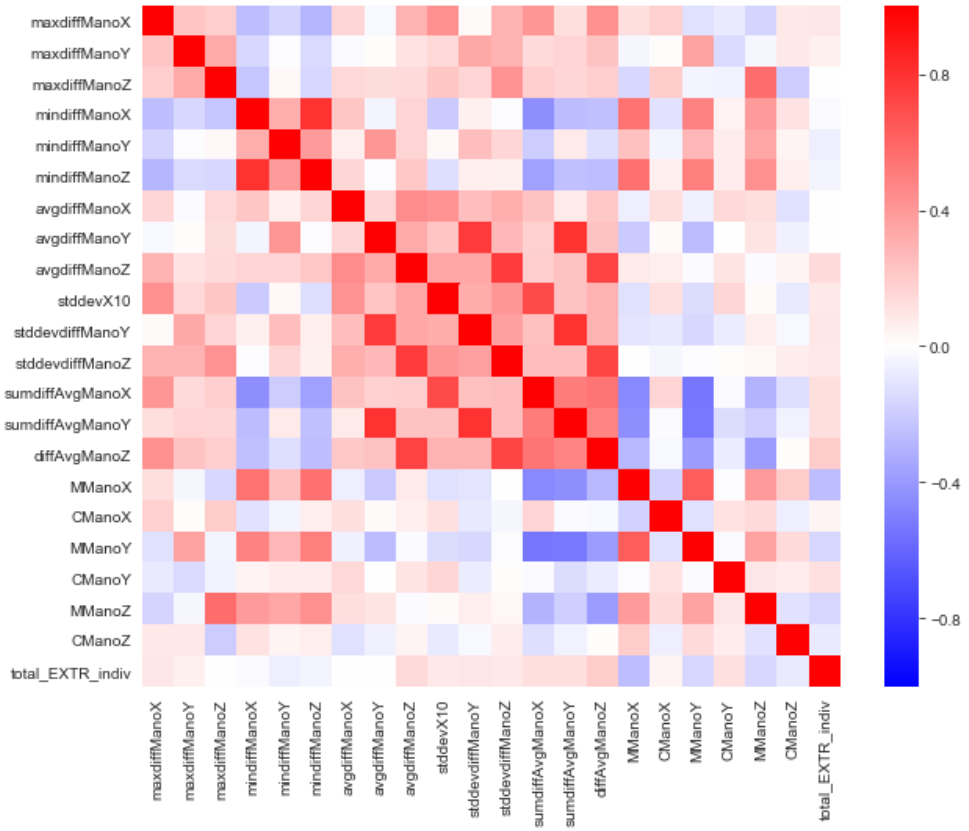


Figura 40. Matriz de correlación – Extraversión individual distancia entre manos

En la Figura 41 se muestra la matriz de correlación correspondiente a la variable extraversión individual, empleando la distancia entre la cabeza y los pies.

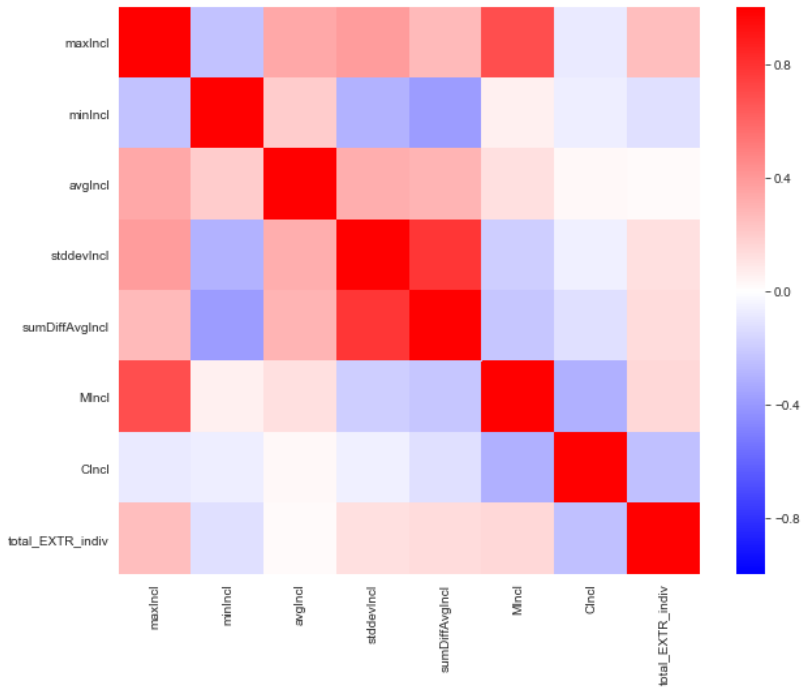


Figura 41. Matriz de correlación – Extraversión individual distancia entre cabeza y pies

En la Figura 42 se muestra la matriz de correlación correspondiente a la variable extraversión individual, empleando la distancia euclídea entre las manos.

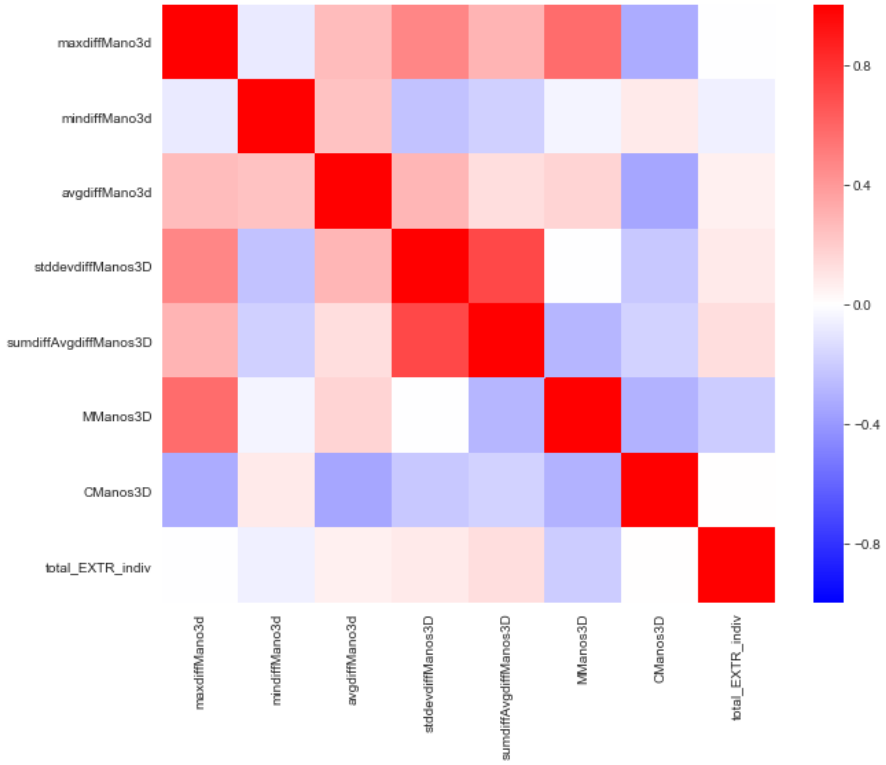


Figura 42. Matriz de correlación – Extraversión individual distancia euclídea entre manos

En la Figura 43 se muestra la matriz de correlación correspondiente a la variable extraversión individual, empleando la distancia entre las manos para el intervalo con mayor movilidad en el eje X para la distancia entre los puntos 8 y 12, y el intervalo con menor movilidad en el eje X para la distancia entre los puntos 8 y 12.

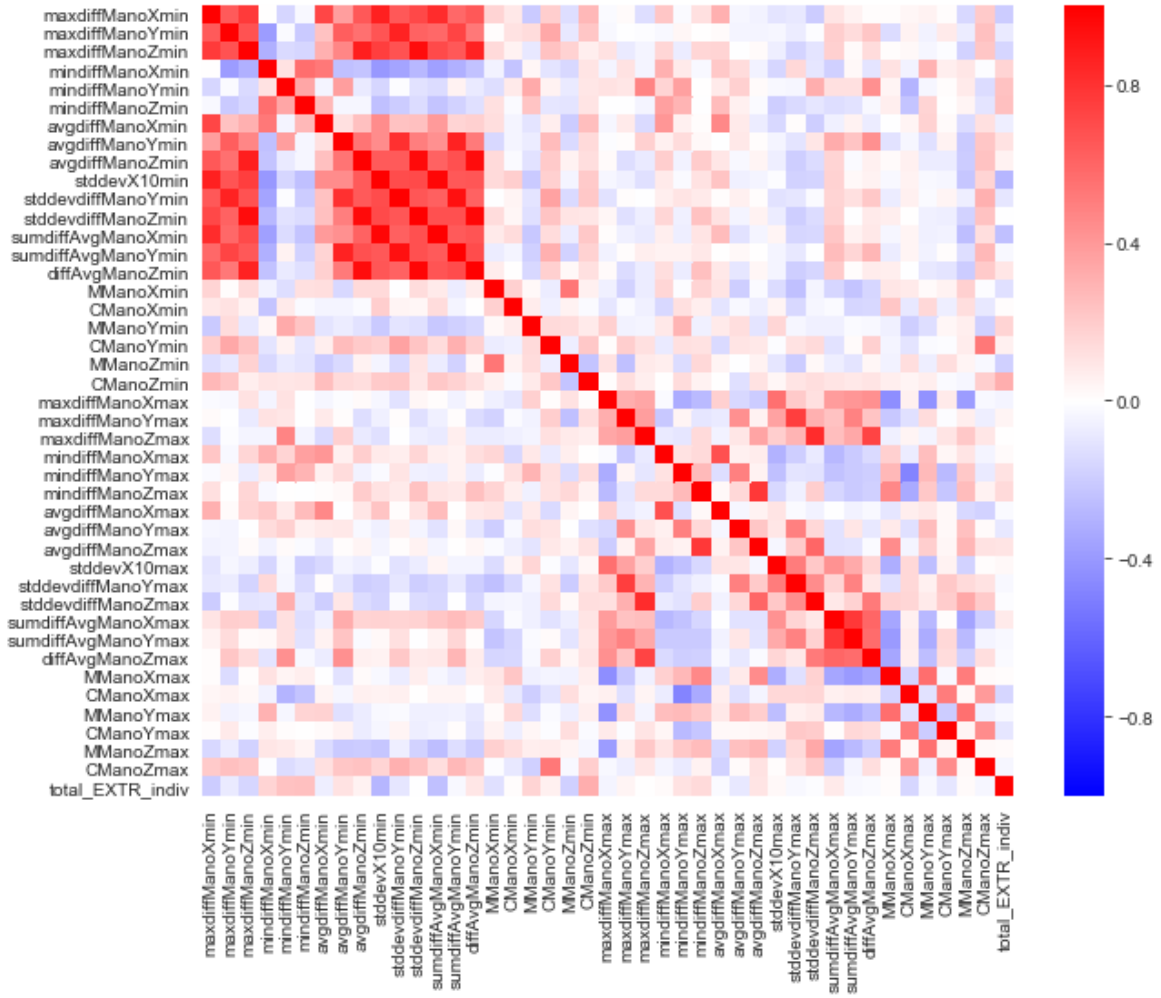
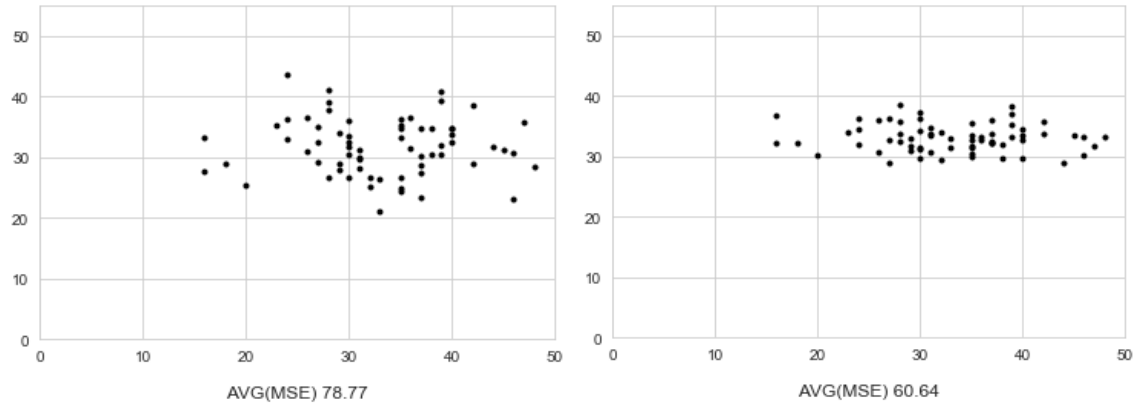


Figura 43. Matriz de correlación – Extraversión individual distancia entre manos para dos intervalos de tiempo

En la Figura 44 se muestra el resultado obtenido para la variable extraversión individual, empleando las características simples mostradas en el Anexo A, con los intervalos de tiempo expuestos anteriormente, y los algoritmos Linear Regression y Random Forest Regression. Como se puede observar, el algoritmo Random Forest proporciona un mejor resultado ya que la media del MSE es menor para este algoritmo, aunque la diferencia entre ambos algoritmos no es muy grande, pero el error obtenido en ambos es grande.



(a) (b)

Figura 44. Regresión Lineal (a) y Random Forest Regression (b) para extraversión individual por intervalos

En la Figura 45 se muestra el histograma para la variable extraversión individual empleando los umbrales descritos en la Tabla 4.

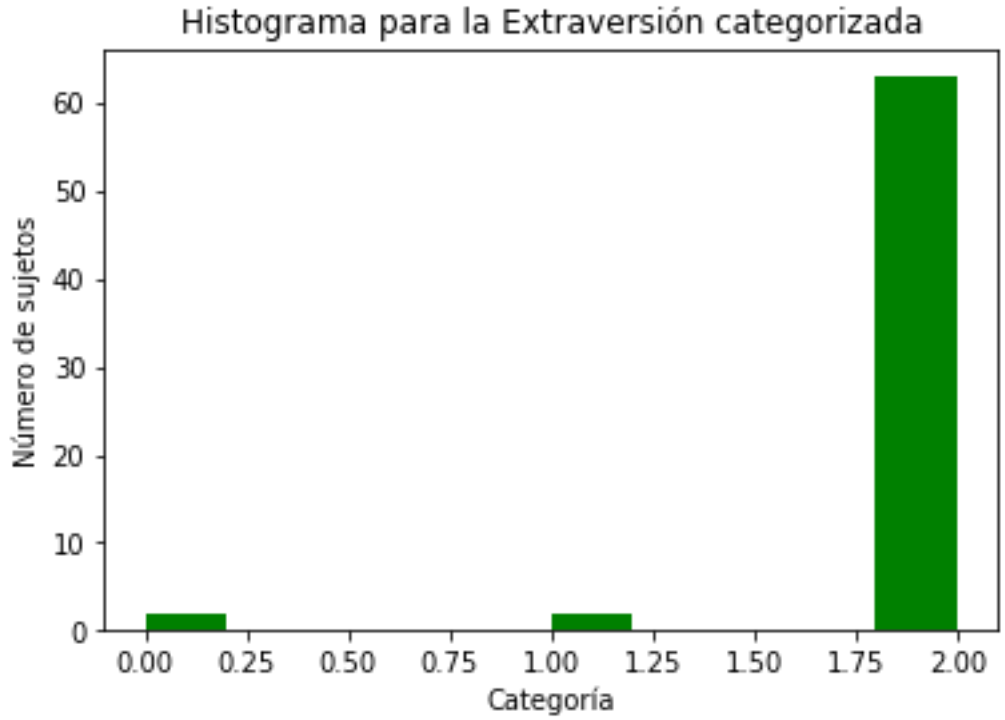


Figura 45. Histograma para la extraversión individual categorizada

En la Figura 46 se muestra el resultado obtenido para la variable extraversión individual, empleando la distancia entre las manos, y el algoritmo Random Forest Classifier. Para ello, se muestra la matriz de confusión, la cual tiene en el eje Y el número de elementos que se correspondían a cada una de las posibles categorías en los datos, y el eje X el número de elementos que se predicen para cada una de las posibles categorías. Además, se puede

observar que el porcentaje de aciertos es bastante elevado, ya que en cada una de las cuatro validaciones cruzadas el número de elementos predichos para cada categoría coincide casi en su totalidad con el número de elementos reales para dicha categoría.

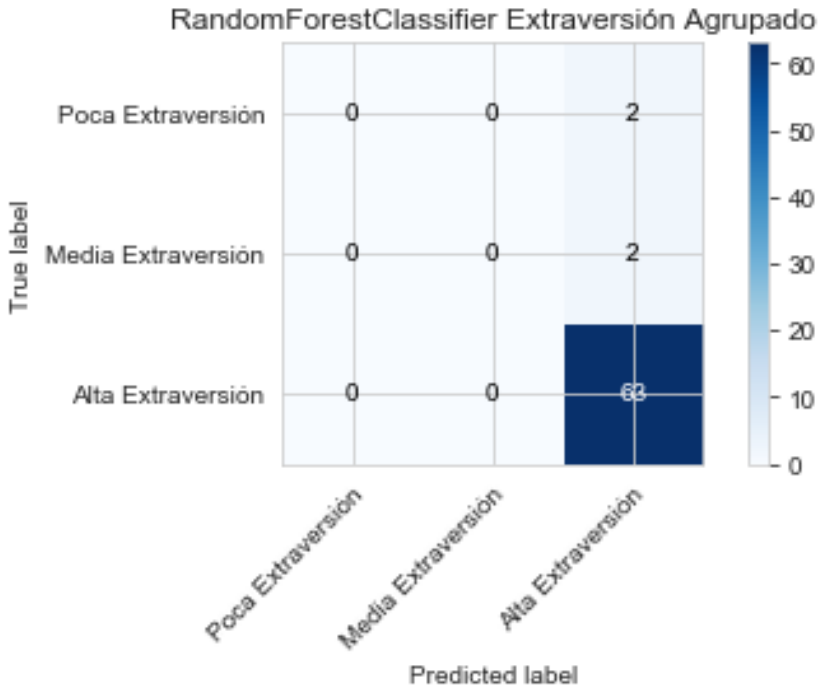


Figura 46. Random Forest Classifier – Extraversión

G Clustering

En este anexo se muestran las diferentes pruebas realizadas para el objetivo de clustering de las variables objetivo.

En la Figura 47 se muestra el resultado para el algoritmo K-means con 4 clusters.

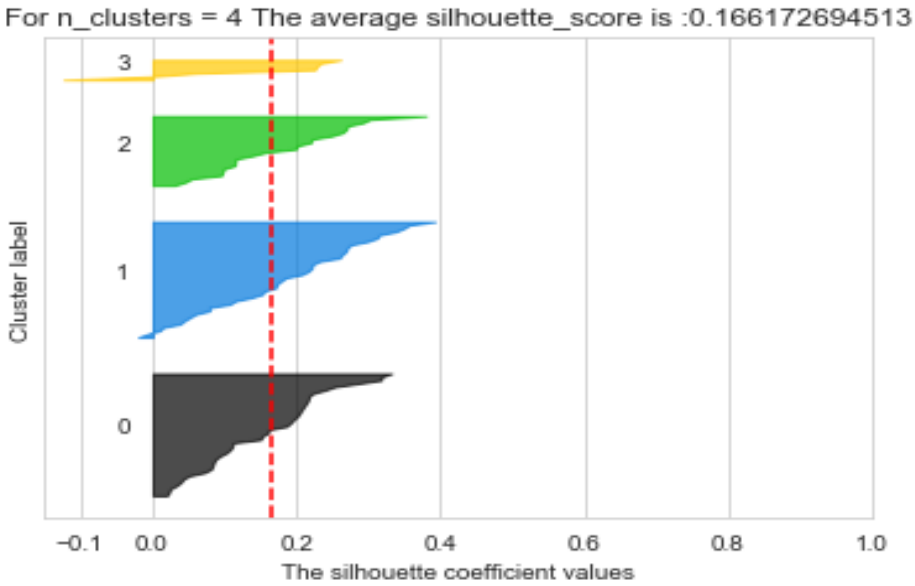


Figura 47. K-means con 4 clusters

En la Figura 48 se muestra el resultado para el algoritmo K-means con 5 clusters.

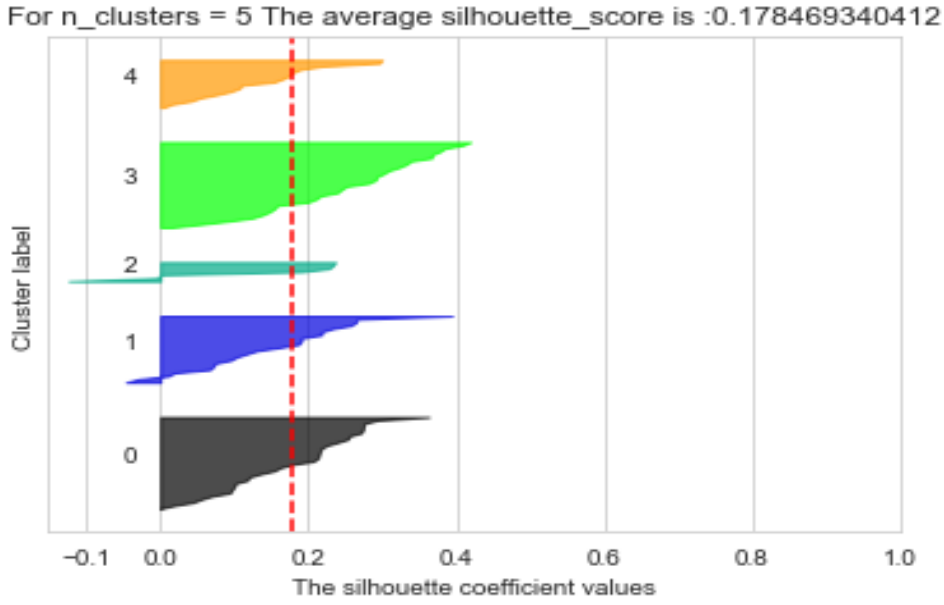


Figura 48. K-means con 5 clusters

En la Figura 49 se muestra un dendograma con los resultados del clustering jerárquico empleando como método single para unir los clusters.

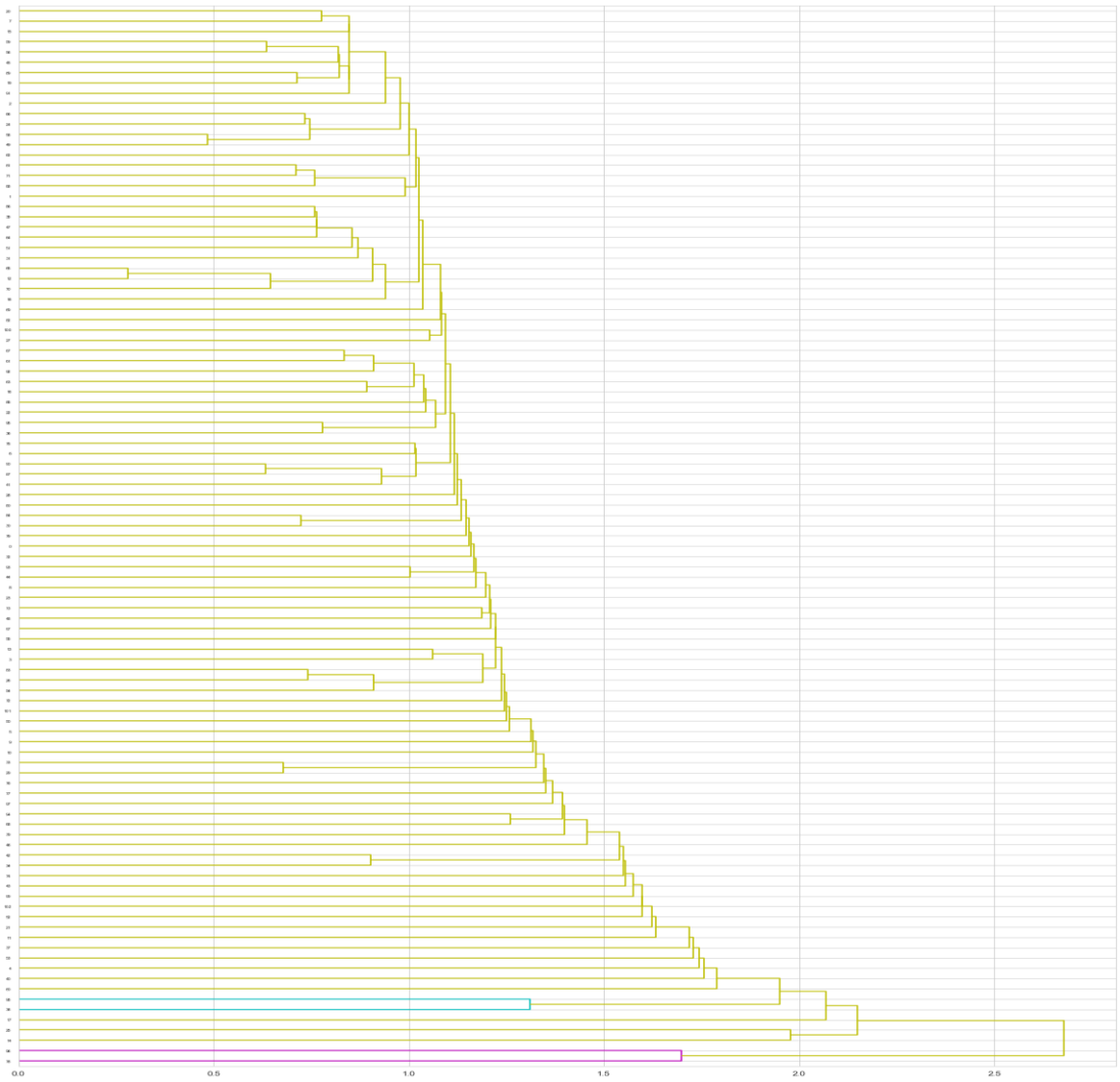


Figura 49. Dendograma clustering jerárquico método single

En la Figura 50 se muestra un dendograma con los resultados del clustering jerárquico empleando como método complete para unir los clusters.

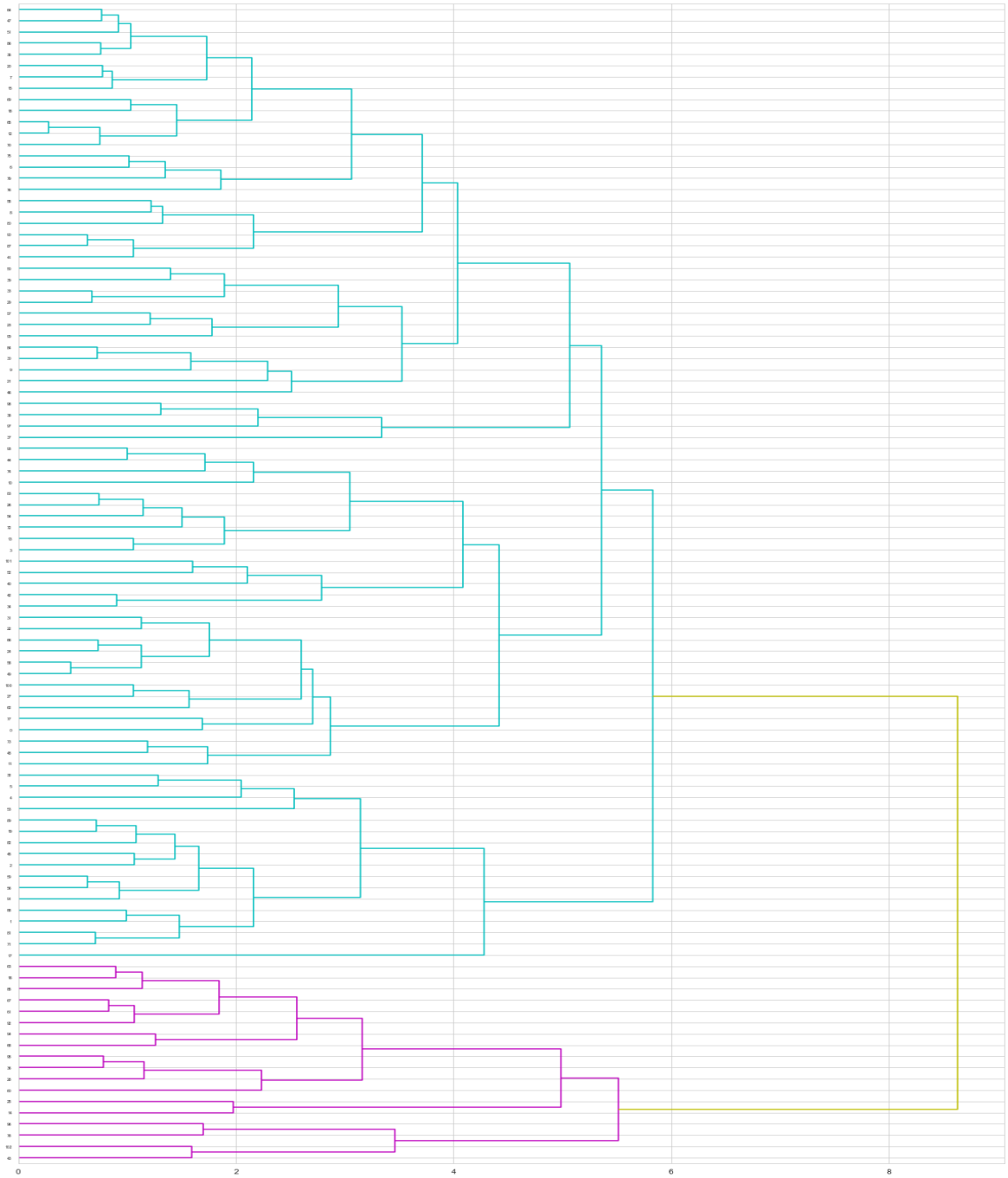


Figura 50. Dendograma clustering jerárquico método complete

En la Figura 51 se muestra un dendograma con los resultados del clustering jerárquico empleando como método average para unir los clusters.

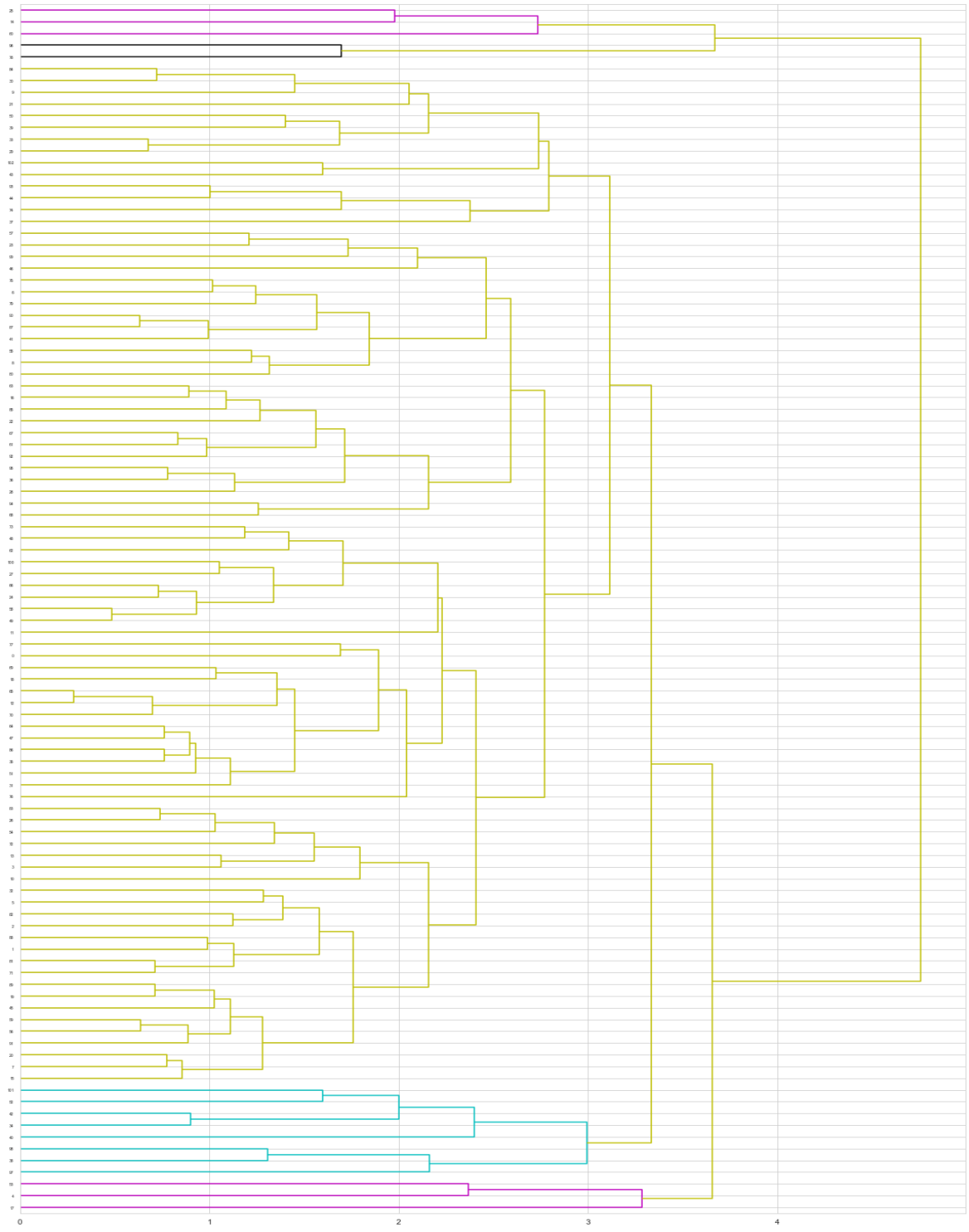


Figura 51. Dendograma clustering jerárquico método average

En la Figura 52 se muestra un dendograma con los resultados del clustering jerárquico empleando como método weighted para unir los clusters.

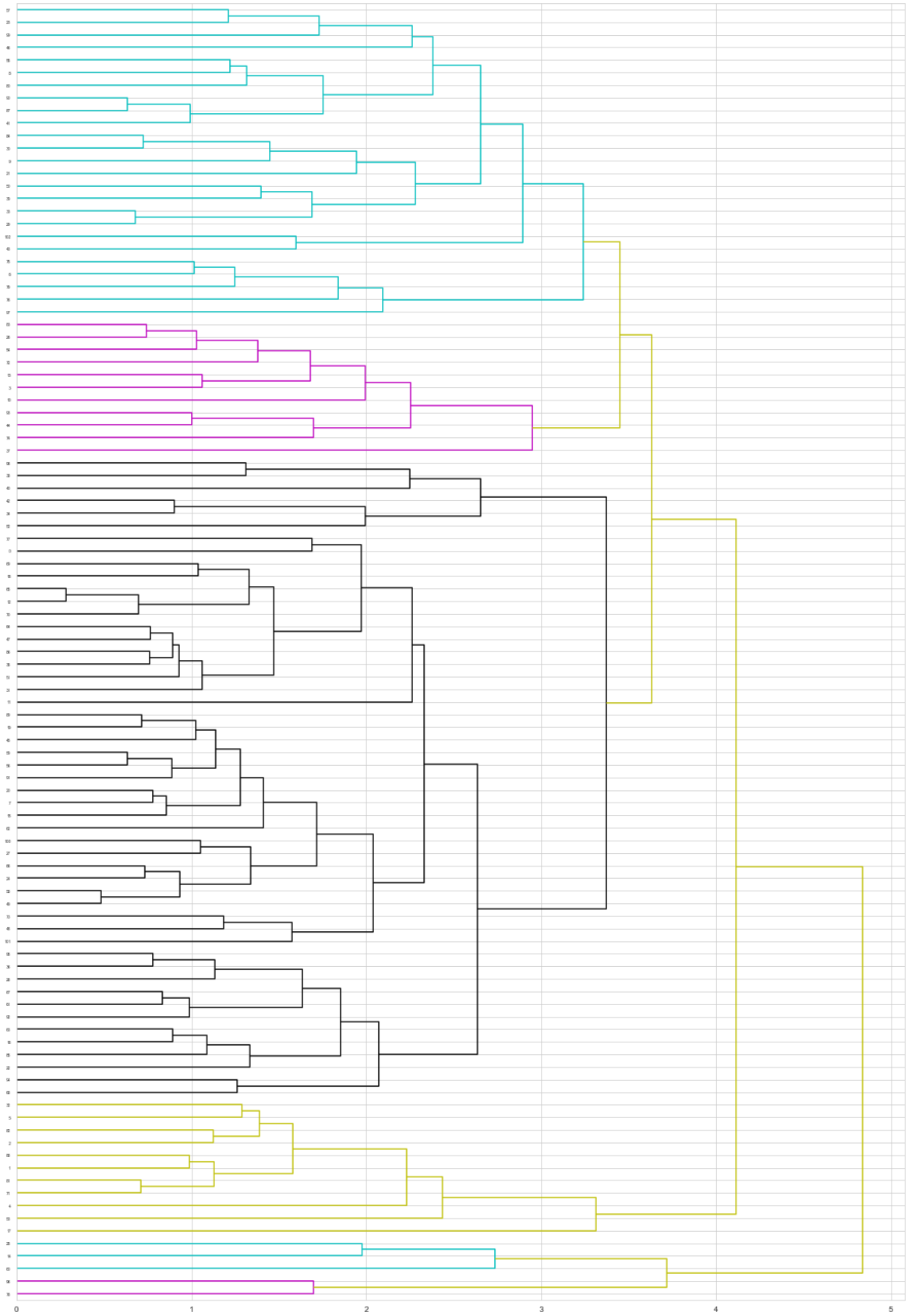


Figura 52. Dendograma clustering jerárquico método weighted

En la Figura 53 se muestra un dendograma con los resultados del clustering jerárquico empleando como método centroid para unir los clusters.

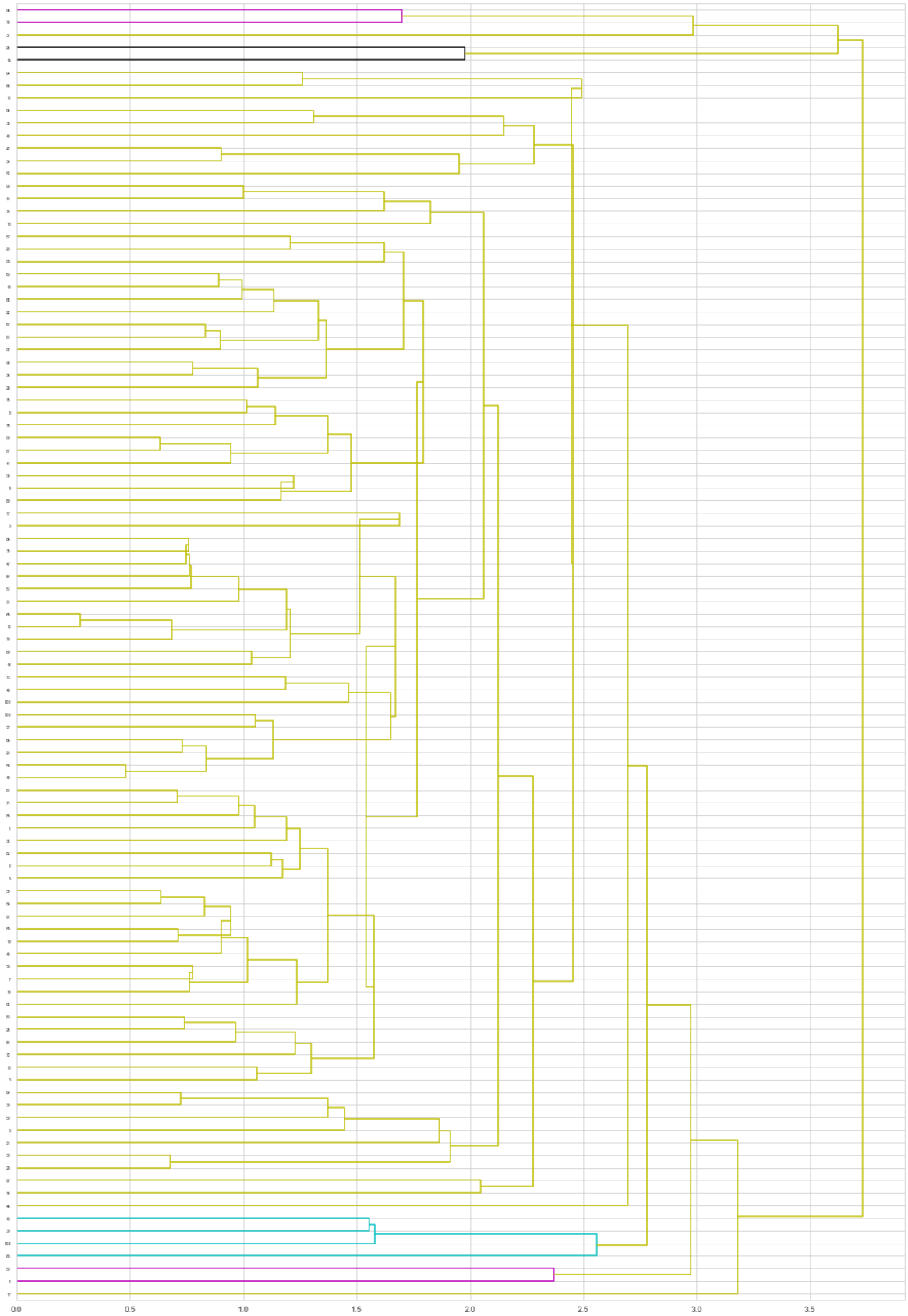


Figura 53. Dendograma clustering jerárquico método centroid

En la Figura 54 se muestra un dendograma con los resultados del clustering jerárquico empleando como método median para unir los clusters.

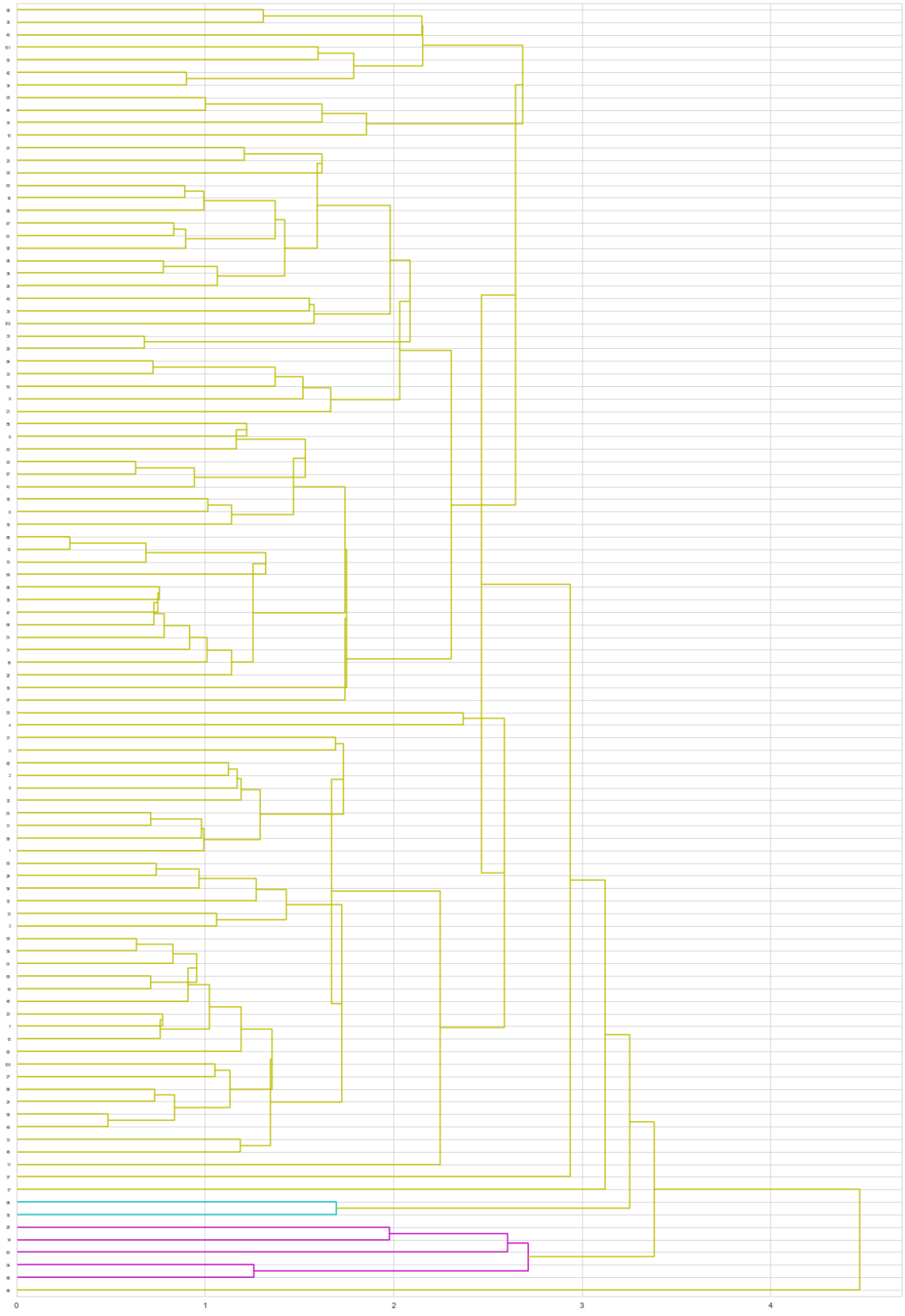
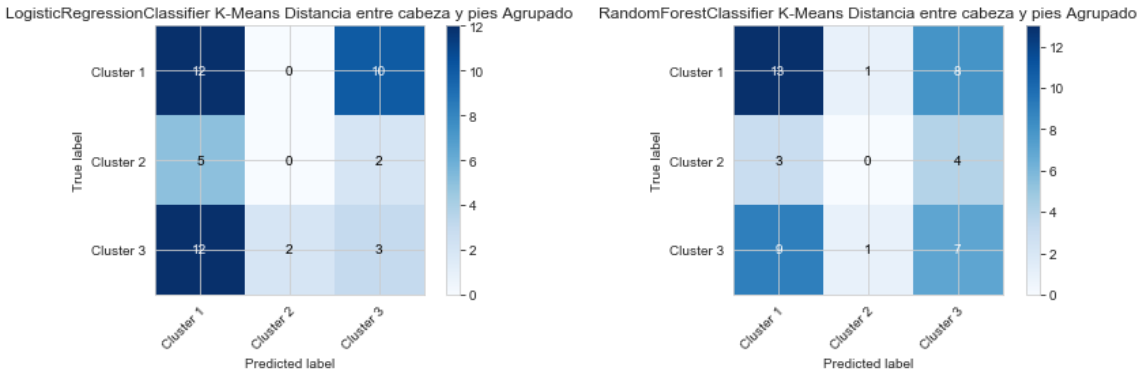


Figura 54. Dendograma clustering jerárquico método median

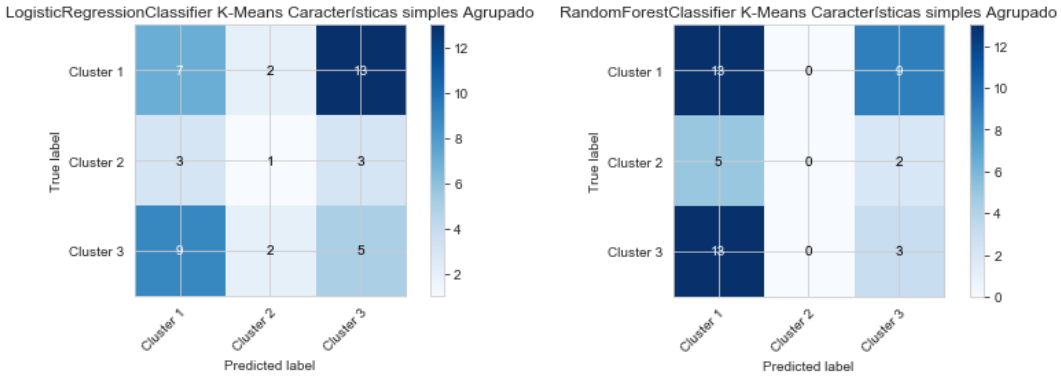
En la Figura 55 se muestran las matrices de confusión para los algoritmos regresión logística y RandomForestClassifier, junto con las clases obtenidas mediante el clustering empleando K-Means con 3 clusters y como características se emplea la distancia entre la cabeza y los pies.



(a) (b)

Figura 55. K-Means - Regresión logística (a) y RandomForestClassifier (b) con distancia entre cabeza y pies

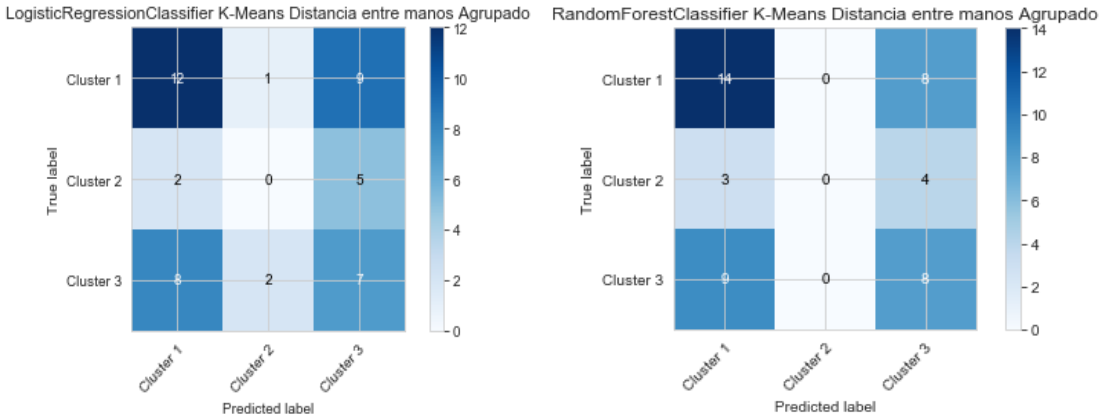
En la Figura 56 se muestran las matrices de confusión para los algoritmos regresión logística y RandomForestClassifier, junto con las clases obtenidas mediante el clustering empleando K-Means con 3 clusters y como características se emplean las características simples.



(a) (b)

Figura 56. K-Means - Regresión logística (a) y RandomForestClassifier (b) con características simples

En la Figura 57 se muestran las matrices de confusión para los algoritmos regresión logística y RandomForestClassifier, junto con las clases obtenidas mediante el clustering empleando K-Means con 3 clusters y como características se emplea la distancia entre las manos.



(a) (b)

Figura 57. K-Means - Regresión logística (a) y RandomForestClassifier (b) con distancia entre manos

En la Figura 58 se muestra la matriz de confusión para el algoritmo RandomForestClassifier, junto con las clases obtenidas mediante el clustering empleando K-Means con 3 clusters y como características se emplea la distancia euclídea entre las manos.

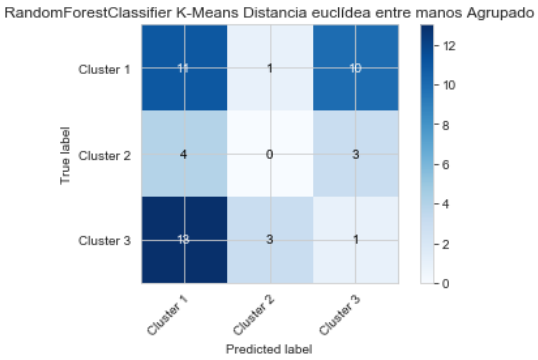
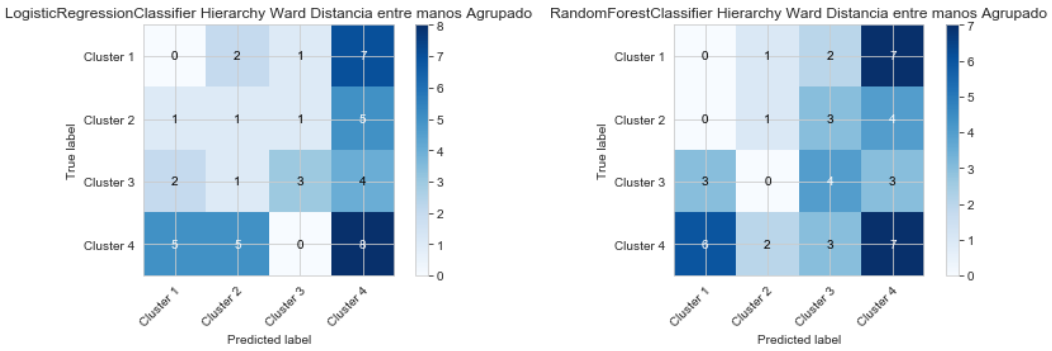


Figura 58. K-Means - RandomForestClassifier con distancia euclídea entre manos

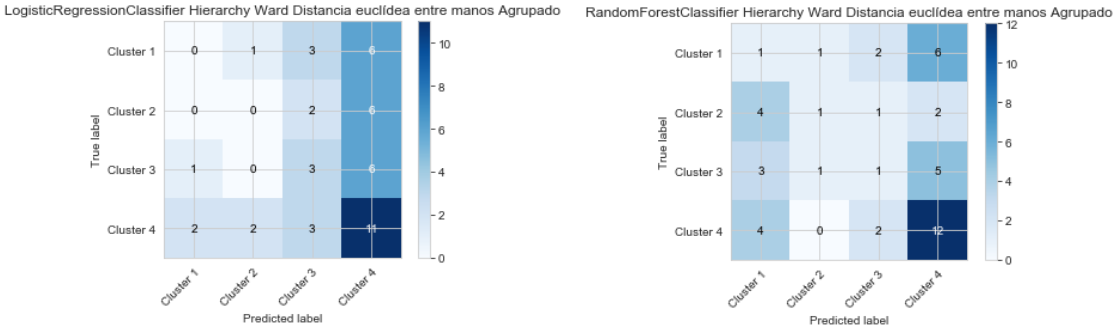
En la Figura 59 se muestran las matrices de confusión para los algoritmos regresión logística y RandomForestClassifier, junto con las clases obtenidas mediante el clustering jerárquico que nos dio como resultado 4 clusters y como característica se emplea la distancia entre las manos.



(a) (b)

Figura 59. Clustering jerárquico - Regresión logística (a) y RandomForestClassifier (b) con distancia entre manos

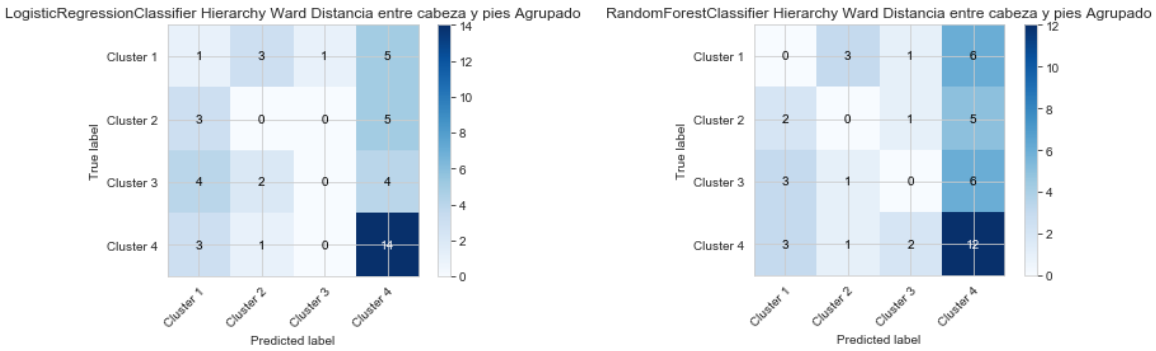
En la Figura 60 se muestran las matrices de confusión para los algoritmos regresión logística y RandomForestClassifier, junto con las clases obtenidas mediante el clustering jerárquico que nos dio como resultado 4 clusters y como característica se emplea la distancia euclídea entre las manos.



(a) (b)

Figura 60. Clustering jerárquico - Regresión logística (a) y RandomForestClassifier (b) con distancia euclídea entre manos

En la Figura 61 se muestran las matrices de confusión para los algoritmos regresión logística y RandomForestClassifier, junto con las clases obtenidas mediante el clustering jerárquico que nos dio como resultado 4 clusters y como característica se emplea la distancia entre la cabeza y los pies.



(a) (b)

Figura 61. Clustering jerárquico - Regresión logística (a) y RandomForestClassifier (b) con distancia entre cabeza y pies

En la Figura 62 se muestra la matriz de confusión para el algoritmo regresión logística junto con las clases obtenidas mediante el clustering jerárquico que nos dio como resultado 4 clusters y como características se emplean las características simples.

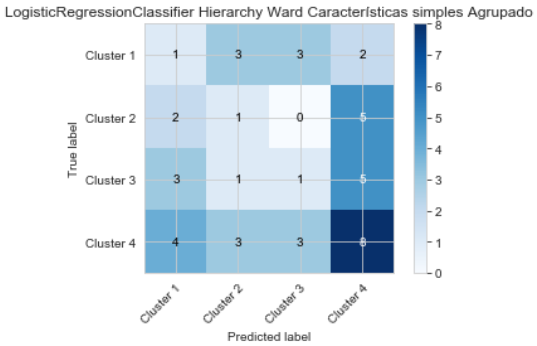


Figura 62. Clustering jerárquico - Regresión logística con características simples